

UTILITY APPLICATION

UNDER 37 CFR § 1.53(B) (2)

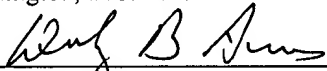
TITLE: METHODS FOR GENETIC ANALYSIS OF DNA  
USING BIASED AMPLIFICATION OF  
POLYMORPHIC SITES

APPLICANT: Vincent P. Stanton, Jr.

Correspondence Enclosed:

Utility Transmittal (2 pgs); Cover Sheet (1 pg); Specification  
(99 pgs); Claims (2 gs); Abstract (1 pg); Drawings (35  
pgs); and Return Postcard

"EXPRESS MAIL" Mailing Label Number EL675944362US Date of Deposit October 25, 2000 I hereby certify under 37 CFR § 1.10 that this correspondence is being deposited with the United States Postal Service as "Express Mail Post Office to Addressee" with sufficient postage on the date indicated above and is addressed to the Commissioner for Patents, Washington, D.C. 20231.

  
\_\_\_\_\_  
Wesley B. Ames

005201-0204560

**DESCRIPTION**

**METHODS FOR GENETIC ANALYSIS OF DNA USING BIASED  
AMPLIFICATION OF POLYMORPHIC SITES**

**RELATED APPLICATION**

This application claims the benefit of Stanton et al., U.S. Provisional Application No. 60/206,613, filed May 23, 2000, entitled METHODS FOR GENETIC ANALYSIS OF DNA, which is hereby incorporated by reference in its entirety, including drawings.

**BACKGROUND OF THE INVENTION**

This application describes methods for the genetic analysis of biologically, medically and economically significant traits in mammals and other organisms, including humans. Genetic analysis refers to the determination of the nucleotide sequence of a gene or genes of interest in a subject organism, including methods for analysis of one site of sequence variation (i.e. genotyping methods) and methods for analysis of a collection of sequence variations (haplotyping methods). Genetic analysis further includes methods for correlating sequence variation with disease risk, diagnosis, prognosis or therapeutic management.

The use of novel genotyping and haplotyping methods for genetic analysis of the apolipoprotein E (ApoE) gene are described. These methods entail use of novel ApoE DNA sequence polymorphisms and haplotypes. The ApoE alleles and genetic analysis methods of this application will allow more sensitive measurement of the contribution of ApoE genetic variation to medically important phenotypes such as risk of heart disease, risk of Alzheimer's disease and response to various therapeutic interventions, including pharmacotherapy.

This application also describes new methods for genotyping a DNA sample based on analysis of the mass of cleaved DNA fragments using mass spectrometry. These genotyping methods are better suited to the present and future requirements of DNA testing than current genotyping methods as a result of improved accuracy, decreased set-up and reagent costs, reduced complexity and excellent compatibility with automation.

At present, DNA diagnostic testing is largely concerned with identification of rare polymorphisms related to Mendelian traits. These tests have been in use for well over a decade. In the future genetic testing will come into much wider clinical and research use, as a means of making predictive, diagnostic, prognostic and pharmacogenetic assessments. These new genetic tests will in many cases involve multigenic conditions, where the correlation of genotype and phenotype is significantly more complex than for Mendelian phenotypes. To produce genetic tests with the requisite accuracy will require new methods that can simultaneously track multiple DNA sequence variations at low cost and high speed, without compromising accuracy. Many

tests will be evaluated in the clinical research setting but only a small fraction will become major diagnostic tests; the clinical research process will reveal that most polymorphisms lack significant functional effects. The genetic analysis methods described in this application are relatively inexpensive to set up and run, while providing extremely high accuracy, and, most important, enabling sophisticated genetic analysis. They are therefore optimally suited to the exigencies of genetic test development in coming years.

The association of specific genotypes with disease risk, prognosis, and diagnosis as well as selection of optimal therapy for disease are some of the benefits expected to ensue from the human genome project. At present, the most common type of genetic study design for testing the association of genotypes with medically important phenotypes is a case control study where allele frequencies are measured in one or more phenotypically defined groups of cases and compared to allele frequencies in controls. (Alternatively, phenotype frequencies in two or more genotypically defined groups are compared.) The majority of such published genetic association studies have focused on measuring the contribution of a single polymorphic site (usually a single nucleotide polymorphism, abbreviated SNP) to variation in a medically important phenotype or phenotypes. In these studies one polymorphism serves as a proxy for all variation in a gene (or even a cluster of adjacent genes).

The limitations of such single polymorphism association analysis are becoming increasingly apparent. Recent articles (e.g. Terwilliger, J. and K.M Weiss. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Current Opinion in Biotechnology* 9: 578-594, 1998) have drawn attention to the low quality of most association studies using single polymorphic sites (evidenced by their low degree of reproducibility). Some of the reasons for the lack of reproducibility of many association studies are apparent. In particular, the extent of human DNA polymorphism – most genes contain 10 or more polymorphic sites, and many genes contain over 100 polymorphic sites - is such that a single polymorphic site can only rarely serve as a reliable proxy for all variation in a gene (which typically covers at least several thousand nucleotides and can extend over 1,000,000 nucleotides). Even in cases where one polymorphic site is responsible for significant biological variation, there is no reliable method for identifying such a site. The haplotyping and genetic analysis methods described in this application provide a systematic way to identify such polymorphic sites.

Several recent studies have begun to outline the extent of human molecular genetic variation. For example, a comprehensive survey of genetic variation in the human lipoprotein lipase (LPL) gene (Nickerson, D. A., et al. *Nature Genetics* 19: 233-240, 1998; Clark, A.G., et al. *American Journal of Human Genetics* 63: 595-612, 1998) compared 71 human subjects and found 88 varying sites in a 9.7 kb region. On average any two versions of the gene differed at 17 sites. This and other studies show that sequence variation may be present at approximately 1 in

100 nucleotides when 50 to 100 unrelated subjects are compared. The implications of the this data are that, in order to create genetic diagnostic tests of sufficient specificity and selectivity to justify widespread medical use, more sophisticated methods are needed for measuring human genetic variation.

Beyond tests that measure the status of a single polymorphic site, the next level of sophistication in genetic testing is to genotype two or more polymorphic sites and keep track of the genotypes at each of the polymorphic sites when calculating the association between genotypes and phenotypes (e.g. using multiple regression methods). However, this approach, while an improvement on the single polymorphism method in terms of considering possible interactions between polymorphisms, is limited in power as the number of polymorphic sites increases. The reason is that the number of genetic subgroups that must be compared increases exponentially as the number of polymorphic sites increases. In a medical study of fixed size this has the effect of dramatically increasing the number of groups that must be compared, while reducing the size of each subgroup to a small number. The consequence of these effects is an unacceptable loss of statistical power. Consider, for example, a clinical study of a gene that contains 10 variable sites. If each site is biallelic then there are  $2^{10} = 1024$  possible combinations of polymorphic sites. If the study population is 500 subjects then it is likely that many genetically defined subgroups will contain only a small number of subjects. Thus, consideration of multiple polymorphisms (as can be determined from DNA sequence data, for example) does not get at the problem that the DNA sequence from a diploid subject does not sufficiently constrain the sequence of the subject's two chromosomes to be very useful for statistical analysis. Only direct determination of the DNA sequence on each chromosome (a haplotype) can constrain the number of genetic variables in each subject to two (allele 1 and allele 2), while accounting for all, or preferably at least a substantial subset of, the polymorphisms.

A much more powerful measure of variation in a DNA segment, then, is a haplotype – that is, the set of polymorphisms that are found on a single chromosome. Because of the evolutionary history of human populations, only a small fraction of all possible haplotypes (given a set of polymorphic sites at a locus) actually occur at appreciable frequency. For example, in a gene with 10 polymorphic sites only a small fraction - perhaps in the range of 1% - of the 1,024 possible genotypes is likely to exist at a frequency greater than 5% in a human population. Further, as described below, haplotypes can be clustered in groups of related sequences to facilitate genetic analysis. Thus determination of haplotypes is a simplifying step in performing a genetic association study (compared to the analysis of multiple polymorphisms), particularly when applied to DNA segments characterized by many polymorphic sites. There is also a potent biological rationale for sorting genes by haplotype, rather than by genotype at one polymorphic

site: polymorphic sites on the same chromosome may interact in a specific way to determine gene function. For example, consider two sites of polymorphism in a gene, both of which encode amino acid changes. The two polymorphic residues may lie in close proximity in three dimensional space (i.e. in the folded structure of the encoded protein). If one of the polymorphic amino acids encoded at each of the two sites has a bulky side chain and the other a small side chain then one can imagine a situation in which proteins that have either [bulky – small], [small – bulky] or [small – small] pairs of polymorphic residues are fully functional, but proteins with [bulky – bulky] residues at the two sites are impaired, on account of a disruptive shape change caused by the interaction of the two bulky side groups. Now consider a subject whose genotype is heterozygous bulky/small at both polymorphic sites. The possible haplotype pairs in such a subject are [bulky – small]/[small – bulky], or [small – small]/[bulky – bulky]. The functional implications of these two haplotype pairs are quite different: active/active or active/inactive, respectively. A genotype test would simply reveal that the subject is doubly heterozygous. Only a haplotype test would reveal the biologically consequential structure of the variation. The interaction of polymorphic sites need not involve amino acid changes, of course, but could also involve virtually any combination of polymorphic sites.

The genetic analysis of complex traits can be made still more powerful by use of schemes to cluster haplotypes into related groups based on parsimony, for example. Templeton and coworkers have demonstrated the power of cladograms for analysis of haplotype data. (Templeton, A.R., Boerwinkle, E. and C.F. Sing. A Cladistic Analysis of Phenotypic Associations With Haplotypes Inferred From Restriction Endonuclease Mapping. I. Basic Theory and an Analysis of Alcohol Dehydrogenase Activity in *Drosophila* *Genetics* 117: 343-351, 1987. Templeton, A.R., Crandall, K.A. and C.F. Sing. A Cladistic Analysis of Phenotypic Associations With Haplotypes Inferred From Restriction Endonuclease Mapping and DNA Sequence Data. *Genetics* 132: 619-633, 1992. Templeton, A.R. and C.F. Sing. A Cladistic Analysis of Phenotypic Associations With Haplotypes Inferred From Restriction Endonuclease Mapping. IV. Nested Analyses with Cladogram Uncertainty and Recombination. *Genetics* 134: 659-669, 1993. Templeton A.R., Clark A.G., Weiss K.M., Nickerson D.A., Boerwinkle E. and C.F. Sing. Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am J Hum Genet.* 66: 69-83, 2000). These analyses describe a set of rules for clustering haplotypes into hierarchical groups based on their presumed evolutionary relatedness. This phylogenetic trees can be constructed using standard software packages for phylogenetic analysis such as PHYLIP or PAUP (Felsenstein, J. Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet.* 22:521-65, 1988; Retief, J.D. Phylogenetic analysis using PHYLIP. *Methods Mol Biol.* 132:243-58, 2000), and hierarchical haplotype clustering can be accomplished using the rules described by Templeton and co-workers. The

methods described by Templeton and colleagues further provide for a nested analysis of variance between different haplotype groups at each level of clustering. The results of this analysis can lead to identification of polymorphic sites responsible for phenotypic variation, or at a minimum narrow the possible phenotypically important sites. Thus, methods for determination of

5 haplotypes have great utility in studies designed to test association between genetic variation and variation in phenotypes of medical interest, such as disease risk and prognosis and response to therapy.

Currently available methods for the experimental determination of haplotypes are unsatisfactory, particularly methods for the determination of haplotypes over long distances (e.g.

10 >5 kb). One of the few experimental haplotyping methods currently in use outside the research group that devised it is based on allele specific amplification using oligonucleotide primers that terminate at polymorphic sites (Newton, C.R. et al. Amplification refractory mutation system for prenatal diagnosis and carrier assessment in cystic fibrosis. *Lancet*. Dec 23-30; 2 (8678-8679):1481-3, 1989; Newton, C.R. et al., Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS) *Nucleic Acids Res.* Vol. 17, 2503-2516, 1989). The method is referred to by the acronym ARMS (for amplification refractory mutation system). The ARMS system was subsequently further developed (Lo, Y.M. et al., Direct haplotype determination by double ARMS: specificity, sensitivity and genetic applications. *Nucleic Acids Research* July 11;19 (13):3561-7, 1991) and has since been used in a number of

20 other studies. ARMS is the subject of US Patents 5,595,890 and 5,853,989. The drawbacks of this method are that (i) the usual limitations of PCR apply in terms of the difficulty of amplifying long DNA segments; (ii) during amplification cycles, an incompletely extended primer extension product may switch (between one or more cycles) from one allelic template strand to the other, resulting in artefactual hybrid haplotypes; (iii) because different DNA samples will be

25 heterozygous at different combinations of nucleotides, different primers and assay conditions for allele specific amplification must be established for each polymorphic site that is to be haplotyped. For example, consider a locus with five polymorphic sites. Subject A is heterozygous at sites 1, 2 and 4; subject B at sites 2 and 3, and subject C at sites 3 and 5. To haplotype A requires allele specific amplification conditions from sites 1 or 4; to haplotype B

30 requires allele specific amplification conditions from sites 2 or 3, and to haplotype C requires allele specific amplification conditions from sites 3 or 5 (with the allele specific primer from site 3 on the opposite strand from that used to haplotype B).

A similar method for achieving allele specific amplification takes advantage of some thermostable polymerases' ability to proofread and remove a mismatch at the 3' end of a primer.

35 Again, primers are designed with the 3' terminal base positioned opposite to the variant base in the template. In this case the 3' base of the primer is modified in a way that prevents it from

being extended by the 5' – 3' polymerase activity of a DNA polymerase. Upon hybridization of the end-blocked primer to the complementary template sequence, the 3' base is either matched or mismatched, depending on which alleles are present in the sample. If the 3' base of the primer is properly base paired the polymerase does not remove it from the primer and thus the blocked 3' end remains intact and the primer can not be extended. However, if there is a mismatch between the 3' end of the primer and the template, then the 3' – 5' proofreading activity of the polymerase removes the blocked base and then the primer can be extended and amplification occurs. This method suffers from the same limitations described above for the ARMS procedure.

Other allele specific PCR amplification methods include further methods in which the 3' terminal primer forms a match with one allele and a mismatch with the other allele (US 5,639,611), PCR amplification and analysis of intron sequences (U.S. 5,612,179 and U.S. 5,789,568), or amplification and identification of polymorphic markers in a chromosomal region of DNA (U.S. 5,851,762). Further, methods for allele-specific reverse transcription and PCR amplification to detect mutations (U.S. 5,804,383), and a primer-specific and mispair extension assay to detect mutations or polymorphisms (PCT/CA99/00733) have been described. Several of these methods are directed to genotyping, not to haplotyping.

Other haplotyping methods that have been described are based on analysis of single sperm cells (Hubert R., Stanton, V.P. Jr, Aburatani H, et al. Sperm typing allows accurate measurement of the recombination fraction between D3S2 and D3S3 on the short arm of human chromosome 3. *Genomics*. 1992 Apr;12(4):683-687); on limiting dilution of a DNA sample (until only one template molecule is present in each test tube, on average) (Ruano, G., Kidd, K.K. and J.C. Stephens. Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. *Proc Natl Acad Sci U S A* 1990 Aug;87(16):6296-6300), or on cloning DNA into various vectors and host microorganisms (US 5,972,614). These methods are not practical for clinical studies of human subjects, and generally have not been used in studies of human disease risk or drug response. For example, sperm based haplotyping methods are not generally useful for clinical studies because no sperm has the same haplotype as its host. Limiting dilution methods are technically challenging – two rounds of PCR amplification are required, with stringent controls for preventing contamination by exogenous DNA -and not compatible with the high throughput, accuracy and reliability required in human clinical studies.

## SUMMARY OF THE INVENTION

This invention concerns methods for determining the sequence of a DNA sample at a polymorphic site, often referred to as genotyping. Many genotyping methods are known in the art, however the methods described in this application have the advantages of being robust, highly accurate, and inexpensive to set up and perform. For these reasons the methods described herein are preferable to currently available methods. The genotyping methods described in the specification may be used in the genotyping steps of the haplotyping methods of this invention, or they may be used for genotyping alone, i.e. not associated with a haplotyping test.

The present invention also concerns methods for determining the organization of DNA sequence polymorphisms on individual chromosomes – i.e. haplotypes, as well as methods for using either genotype or haplotype information, or a combination of the two, to make diagnostic tests useful for disease risk assessment, for prognostic prediction of the course or outcome of a disease, to diagnose a disease or condition, or to select optimal therapy for a disease or condition. As described above, haplotypes are often not directly inferrable from genotypes, therefore specialized methods are required to determine haplotypes. Further, as noted, currently available haplotyping methods are cumbersome and/or are limited by the type of samples that can be analyzed. The several haplotyping methods of this invention are superior to previously described methods with respect to technical ease, sample throughput, length of DNA that can be haplotyped, and compatibility with automation. These novel methods provide the basis for more sophisticated analyses of the contribution of variation at candidate genes (such as ApoE) to intersubject variation in medical or other phenotypes of interest. These methods are applicable to patients with a disease or disorder as well as to apparently normal subjects in whom a predisposition to a disease or disorder may be discovered or quantified as a result of a haplotyping test described herein. Application of the haplotyping methods of this invention will provide for improved medical care by increasing the accuracy of genetic diagnostic tests of all kinds.

This invention further concerns genetic analysis of the Apo E gene to determine disease and drug response traits in humans, particularly traits that may be affected by genetic variation at the ApoE gene, and further concerns methods for improving medical care for individual patients based on the results of ApoE genetic testing. Variation at the ApoE gene has been associated with risk of Alzheimer's disease and other neurodegenerative diseases, recovery from organic or traumatic brain injury, and response to pharmacotherapy of AD as well as coronary heart disease, dyslipidemia, and other conditions. The methods of this application also provide for more efficient use of medical resources, and therefore are also of use to organizations that pay for health care, such as managed care organizations, health insurance companies and the federal government. The invention provides methods for performing genotyping and haplotyping tests



on a human subject to formulate or assist in the formulation of a diagnosis, a prognosis or the selection of an optimal treatment method based on ApoE genotype or haplotype. These methods are applicable to patients with a disease or disorder affecting the cardiovascular or nervous systems, as well as patients with any disease or disorder that is affected by lipid metabolism. The ApoE haplotyping methods of this invention are equally applicable to apparently normal subjects in whom predisposition to a disease or disorder may be discovered as a result of an ApoE genotyping or haplotyping test described herein. Application of the methods of this invention will provide for improved medical care by, for example, allowing early implementation of preventive measures in patients at risk of diseases such as atherosclerosis, dementia, Parkinson's disease, Huntington's disease or other organic or vascular neurodegenerative process; or optimal selection of therapy for patients with diseases or conditions such as hyperlipidemia, cardiovascular disease (including coronary heart disease as well as peripheral or central nervous system atherosclerosis), neurological diseases including but not limited to Alzheimer's disease, stroke, head or brain trauma, amyotrophic lateral sclerosis, and psychiatric diseases such as psychosis, bipolar disease and depression.

### Genotyping Methods

The disadvantages of existing genotyping methods include unproven or inadequate accuracy (particularly for medical research or clinical practice, where very high accuracy is required), high set up costs (which are unacceptable when relatively small numbers of subjects are being studied – e.g. in the clinical research setting), technical difficulty in performing the test or interpreting the results, and incompatibility with full automation.

Methods described in the present invention first use amplification (preferably PCR amplification) using amplification oligonucleotides (primers) flanking a polymorphic site. The 3' end of one of the primers is close, highly preferably within 16 nucleotides, of a polymorphic site in template DNA. The second primer may lie at any distance from the first primer on the opposite side of the polymorphic site providing effective amplification. The first primer is designed so that it introduces two restriction endonuclease recognition sites into the amplified product during the amplification process. Preferably the two restriction sites are created by inserting a sequence of 15 or fewer nucleotides into the primer. This short inserted sequence in general does not base pair to the template strand, but rather loops out when the primer is bound to template. However, when the complementary strand is copied by polymerase the inserted sequence is incorporated into the amplicon. Incubation of the resulting amplification product with the appropriate restriction endonucleases results in the excision of a small (generally < 20 bases) polynucleotide fragment that contains the polymorphic nucleotide. The small size of the excised fragment allows it to be easily and robustly analyzed by mass spectrometry to determine

the identity of the base at the polymorphic site. The primer with the restriction sites can be designed so that the restriction enzymes: (i) are easy to produce, or inexpensive to obtain commercially, (ii) cleave efficiently in the same buffer, i.e. all potential cleavable amplicons are fully cleaved in one step, (iii) cleave multiple different amplicons, so as to facilitate multiplex analysis (that is, the analysis of two or more samples simultaneously).

An enhancement of the basic method is to select a combination of restriction enzymes that will cleave the amplified product so as to produce staggered ends with a 5' extension, such that the polymorphic site is contained in the extension. Elimination of natural nucleotides from the reaction (for example using Shrimp Alkaline Phosphatase or other alkaline phosphatase) and addition of at least one modified nucleotide corresponding to one of the two nucleotides present at the polymorphic site (for example 5'-bromodeoxyuridine if T is one of the two polymorphic nucleotides) will result in fill-in of the recessed 3' end to produce fragments differing in mass by more than the natural mass difference of the two polymorphic nucleotides. One or more modified nucleotides can be selected to maximize the differential mass of the two allelic fill-in products. This enhancement of the basic method has the advantage of reducing the mass spectrometric resolution required to reliably determine the presence of two alleles vs. one allele, thereby improving the performance of base-calling software and the ease with which a genotyping system can be automated.

Another modification of the basic system is to use a third restriction enzyme that cleaves only one of the two alleles, such that the presence of the site yields shorter fragments than are observed in its absence. Such a modification is not universally applicable because not all polymorphisms alter restriction sites, however this limitation can be partially addressed by including part of the restriction enzyme recognition site in the primer. For example, an interrupted pallindrome recognition site like Mwo I (GCNNNNN/NNGC) can be positioned such that the first GC is in the primer while the second GC includes the polymorphic nucleotide. Only the allele corresponding to GC at the second site will be cleaved. Use of such restriction endonucleases simplifies the sequence requirements at and about the polymorphic site (in this example all that is required is that one allele at the polymorphic site include the dinucleotide GC), thereby increasing the number of polymorphic sites that can be analyzed in this way.

In additional aspects, the invention provides methods that are applicable to both genotyping and haplotyping. The methods use biased amplification of nucleic acid sequences that include variance sites, and utilize primers that are designed so that a hairpin loop will form, generally in the complementary strand formed in an amplification reaction. The primer is designed to have a mismatch in its 5' end to a particular nucleotide at a particular site, generally a polymorphic site in a gene. If the particular nucleotide is present at the site, then amplification will be inhibited because the complementary strand formed in the amplification reaction will

form a sufficiently stable hairpin loop to effectively compete with binding of the primer, and so inhibit further amplification. In contrast, a variant sequence with a different nucleotide at that site will not form a sufficiently stable hairpin to effectively compete with primer binding.

Thus, in one aspect, the invention provides a method for biasing the amplification of one allele (e.g., one form of a SNP at a particular site). As explained above, the biasing depends on the identity of a specific nucleotide at a polymorphic site in a target nucleic acid sample. The method involves contacting a segment of DNA with two primers encompassing the polymorphic site under amplification conditions. One primer contains a region at its 5' end that is not complementary to the target nucleic acid but which, when incorporated into the amplification product, will cause the 3' end of the strand complementary to this primer in the amplification product to form a sufficiently stable hairpin loop by hybridizing with the sequence including the polymorphic site to inhibit further amplification only if the specific nucleotide is present at the polymorphic site. The method also involves determining whether the segment is amplified. Amplification (or preferential amplification) of the segment is indicative that the polymorphic site contains an alternative to the specific nucleotide.

In particular embodiments, the nucleic acid sample can be single stranded DNA or double stranded DNA, and can be genomic or cDNA. RNA can also be utilized, preferably by forming cDNA.

In certain embodiments, the amplification of the segment is detected by detection of the presence of defined size fragments following restriction enzyme digestion of any amplification products. The polymorphic site can be a restriction fragment length polymorphism (RFLP), and a digestion can be performed with a restriction enzyme corresponding to the RFLP, where the defined size fragments differ in size depending on the nucleotide present at the polymorphic site.

The method is not restricted to a single site, so in preferred embodiments, the method involves carrying out the contacting and determining for each of a plurality of different polymorphic sites. For example, at least 2, 3, 4, 5, 6, 8, 10, 15, 20, 30, 40, 50, or 100 sites can be analyzed in a coordinated set of determinations (e.g., in genotyping an individual for a plurality of different sites, which may be in one or a plurality of different genes). In certain embodiments, the plurality of different polymorphic sites provides a haplotype for a gene, can independently or also include at least one polymorphic site in a plurality of different genes, and/or provide haplotypes for a plurality of different genes.

Such biased amplification can be used to determine the nucleotide present at a particular polymorphic site. Thus, in a related aspect, the invention provides a method for determining whether a particular nucleotide is present at a polymorphic site in a target nucleic acid sequence, by contacting a segment of DNA containing the polymorphic site with a primer under amplification conditions, such that extension products and/or amplification products will be

formed. The primer has a sequence at its 5' end that is the same as a sequence including the polymorphic site for a particular nucleotide present at that site. The opposite strand extension product or amplification product will form a sufficiently stable hairpin loop by hybridization between a sequence including the polymorphic site and a sequence derived from the 5' end of the primer for a specific nucleotide at the polymorphic site to inhibit amplification.

Amplification is not inhibited for an alternative nucleotide at said site. The method also includes determining whether the segment is amplified. Amplification of the segment indicates that the polymorphic site contains an alternative nucleotide instead of the specific nucleotide. In general, a second primer, constituting a primer pair, is also used under amplification conditions such that extension products or amplification products or both will be formed. Particular embodiments include those as described for the aspect above.

### Haplotyping Methods

This invention concerns methods for determining the sequence of individual chromosomes, starting with diploid DNA that contains two chromosomes, and methods for using that information to make genetic tests useful for disease risk assessment, for diagnosing a disease or condition, for assessing disease prognosis or to select optimal therapy for a disease or condition. The sequence of a chromosome segment is referred to as a haplotype. Since homologous chromosome segments (e.g. the sequence of two alleles of the ApoE gene) are very similar in sequence (>99%) the distinguishing elements of haplotypes occur at polymorphic sites. A haplotype can be thought of as the nucleotide sequence of a DNA segment at some or all of the sites that vary in a population. Thus a haplotype may consist in specifying the sequence at 10 polymorphic sites in a 5,000 nucleotide DNA segment.

The pattern of genetic variation in most species, including man, is not random; as a result of human evolutionary history some sets of polymorphisms occur together on chromosomes, so that knowing the sequence of one polymorphic site may allow one to predict with some probability the sequence of certain other sites on the same chromosome. Once the relationships between a set of polymorphic sites have been worked out, a subset of all the polymorphic sites may be used in the development of a haplotyping test. In preferred embodiments of the haplotyping methods of this invention, a subset of all the polymorphic sites at a locus is used to develop a haplotyping test. The polymorphisms that comprise a haplotype may be of any type.

Most polymorphisms (about 90% of all DNA polymorphisms) involve the substitution of one nucleotide for another, and are referred to as single nucleotide polymorphisms (SNPs). The other main type of polymorphism involves change in the length of a DNA segment as a result of an insertion or deletion of anywhere from one nucleotide to thousands of nucleotides.

Insertion/deletion polymorphisms (also referred to as indels) account for most non-SNP

polymorphisms. Common kinds of indels include variation in the length of homopolymeric sequences (e.g. AAAAAA vs. AAAAA), variation in the number of short tandem repeat sequences such as CA (e.g. 13 repeats of CA vs. 15 repeats), and variation in the number of more complex repeated sequences (sometimes referred to as VNTR polymorphisms, for variable number of tandem repeats), as well as any other type of inter-individual variation in the length of a given DNA segment. The repeat units may also vary in sequence.

Haplotypes are often not directly inferable from genotypes (except in the special case of families, where haplotypes can often be inferred by analysis of pedigrees), therefore specialized methods are required for determining haplotypes from samples derived from unrelated subjects:

Currently available haplotyping methods are cumbersome and expensive and limited either by the type of samples that can be analyzed (e.g. sperm cells) or by the limitations of PCR or other DNA amplification methods. The limits of DNA amplification methods such as PCR include incomplete allele-specificity of priming when using a 3' terminal primer mismatch to achieve allele discrimination (such as in the ARMS method); that is, there may be some amplification of the non-selected allele. PCR is also limited in the length of DNA segment that can be amplified.

The present application provides methods for determining the haplotypes present in a DNA sample or cDNA sample preferably drawn from one subject, however these methods may also be used to determine the population of haplotypes present in a complex mixture, such as may be produced by mixing DNA samples from multiple subjects. The methods described herein are applicable to genetic analysis of any diploid organism, or any polyploid organism in which there are only two unique alleles. Application of the methods of this invention will provide for improved genetic analysis, enabling advances in medicine, agriculture and animal breeding. For example, by improving the accuracy of genetic tests for diagnosing predisposition to disease, or for predicting response to medical therapy, it will be possible to make safer and more efficient use of appropriate preventive or therapeutic measures in patients. The methods of this invention also provide for improved genetic analysis in a variety of basic research problems, including the identification of alleles of human genes that are associated with disease risk or disease prognosis.

Certain methods for determining haplotypes present in a DNA sample from a diploid organism include the following steps: (i) genotyping at least a portion of (meaning a sequence portion) the sample to identify sites of heterozygosity; (ii) enriching for an allele by a method not requiring amplification to a ratio of at least 1.5:1 based on a starting ratio of 1:1, where the information from (i) is used to select a preferred or optimal heterozygous site or sites for allele enrichment; (iii) genotyping the enriched material to determine the nucleotides present at said heterozygous site or sites; and (iv) determining the haplotype of the enriched allele by inspecting the genotypes from (iii). This method may further include determining the haplotype of the non-

enriched allele by comparing the genotype determined in step (i) of with the haplotype determined in step (iv). Such a haplotyping method as described above may include additional steps including (a) performing an allele enrichment procedure for the second allele on the same starting material and (b) genotyping the enriched material for the second allele to determine the nucleotides present at said heterozygous site or sites; and (c) determining the haplotype of the enriched second allele by inspecting the genotypes from (b).

Additional methods for determining the haplotypes present in DNA from a diploid organism, include the following steps: (i) genotyping at least a portion of the DNA in a sample from said organism to identify sites of heterozygosity; (ii) performing an allele-selective amplification procedure on the sample such that the allele ratio is changed from a starting ratio of 1:1 to at least 1.5:1, wherein the information from (i) is used to select an optimal polymorphic site or sites for designing primers to achieve said allele-selective amplification; (iii) genotyping the selectively amplified material; and (iv) determining the haplotype of the selectively amplified allele by inspecting the genotypes. Methods may include further determination of the haplotype of the selectively non-amplified allele by comparing the genotype determined in (i) with the haplotype determined in (iv). In addition, methods may include determining the haplotype of the selectively non-amplified allele by (a) performing an allele-selective amplification procedure for the second allele using the same starting material; (b) genotyping the selectively amplified second allele material; and (c) determining the haplotype of the selectively amplified second allele by inspecting the genotypes.

Also, methods for determining the haplotypes present in DNA from a diploid organism, include (i) genotyping at least a portion of a DNA sample from said organism to identify sites of heterozygosity that affect restriction enzyme cleavage sites; (ii) restriction endonuclease digesting the DNA, using natural or synthetic endonucleases, such that one allele is restricted at a specific site and the other is not; (iii) performing an amplification procedure on the sample, using the information from step (i) to select optimal sites for designing primers to achieve allele-selective amplification; (iii) genotyping the selectively amplified material; and (iv) determining the haplotype of the selectively amplified allele by inspecting the genotypes. These haplotyping methods further include determining the haplotype of the selectively non-amplified allele by comparing the genotype determined in step (i) with the haplotype determined in step (iv). In addition, methods may include (a) isolating the second allele utilizing size difference; (b) genotyping the size selected material corresponding to the second allele; and (c) determining the haplotype of the size-selected second allele by inspecting the genotypes.

Still further methods for determining the haplotypes present in DNA from a diploid organism include the steps of (i) genotyping at least a portion of the DNA from the sample to identify sites of heterozygosity that affect restriction enzyme cleavage sites; (ii) restriction endonuclease digesting the DNA, using natural or synthetic endonucleases, such that only one allele is restricted at a specific polymorphic site, thereby creating partially overlapping allele 1 and allele 2 fragments of different length, wherein information from (i) is utilized to select a restriction site that produces a useful difference in allele length; (iii) separating the restricted molecules according to their size by electrophoresis or centrifugation, such that the two allelic restriction fragments are resolved; isolate DNA molecules corresponding to the size of allele 1 and, optionally, allele 2; (iv) genotyping the size selected material corresponding to allele 1 and optionally allele 2; and (v) determining the haplotype of the size-selected allele 1 by inspecting the genotypes. These methods may include determination of the haplotype of allele 2 by comparing the genotypes determined in (i) with the haplotype determined in (v).

Additional embodiments of methods for haplotyping double stranded DNA fragments include (i) genotyping at least a portion of a DNA sample to identify sites of heterozygosity in the DNA fragment of interest; (ii) immobilizing double stranded DNA fragments on a solid support ; (iii) adding two or more components that bind at polymorphic sites in the immobilized DNA fragment of interest to produce detectable structure under conditions that promote preferential binding to only one strand of the target immobilized fragment; and (iv) determining the location of target fragments. These methods may further include two or more components which are two or more oligonucleotides complementary to polymorphic sites in the aforementioned immobilized DNA fragment of interest. The components are added under conditions that promote D loop formation in the case of oligonucleotides perfectly matched to one strand of the target immobilized fragment, but not in the case of oligonucleotides containing one or more mismatched nucleotides. The formation of D loops may be enhanced by the addition of RecA protein or alternatively by the alteration of salt concentration within the mixture. The two or more components may further include two or more peptide nucleic acids (PNA) or two or more zinc finger proteins. In methods including PNA, the peptide nucleic acids are complementary to polymorphic sites in the immobilized DNA fragment of interest, and are added under conditions that promote D loop formation in the case of PNAs perfectly matched to one strand of the target immobilized fragment, but not in the case of PNAs containing one or more mismatched nucleotides. In methods including zinc finger proteins, the proteins that can bind to one of two alleles at a polymorphic nucleotide may be used and are added as described for the oligonucleotide components. The two or more zinc finger proteins can be detectably labeled. The immobilized target DNA fragments may be first subjected to a size selection

procedure and or immobilized to a prepared glass surface. These methods may then be used to determine the location of the target fragments by optical mapping. In this more specific method for detection, two or more oligonucleotides are detectably labeled.

Further embodiments of a method for determining the haplotypes of DNA fragments present in a DNA sample from a diploid organism including: a) selectively amplifying one haplotype from the mixture by the allele specific clamp PCR procedure; and b) determining the genotype of two or more polymorphic sites in the amplified DNA fragment. The selective amplification may be preceded by determining the genotype of the DNA sample at two or more polymorphic sites in order to devise an optimal genotyping and that the DNA sample is a mixture of several DNA samples.

Additional haplotyping methods and embodiments of this invention are described in the Detailed Description below.

#### APOE genotyping and haplotyping

Several United States patents relate to methods for determining ApoE haplotype and using that information to predict whether a patient is likely to develop late onset type Alzheimer's Disease (US Patents 5508167, 5716828), whether a patient with cognitive impairment is likely to respond to a cholinomimetic drug (US Patent 5935781), or whether a patient with a non-Alzheimer's neurological disease is likely to respond to therapy (US Patent 5508167).

The ApoE test practiced in all the cited patents (and virtually all the other publications), is based on a classification of Apo E into three alleles, termed epsilon 2, epsilon 3 and epsilon 4 (and abbreviated e2, e3 and e4). These three alleles are distinguishable on the basis of two polymorphic sites in the ApoE gene. The status of both sites must be tested to determine the alleles present in a subject. The two polymorphic sites are at nucleotides 448 and 586 of the ApoE cDNA (numbering from GenBank accession K00396), corresponding to amino acids 112 and 158 of the processed ApoE protein. The nucleotide polymorphism at both sites is T vs. C, and at both sites it is associated with a cysteine vs. arginine amino acid polymorphism, wherein the codon with T encodes cysteine and the codon with C encodes arginine. The presence of T at both polymorphic sites (cysteine at both residues 112 and 158) is designated e2; T at position 448 and C at position 586 (cysteine at 112, arginine at 158) is designated e3, and C at both variable sites (arginine at both 112 and 158) is designated e4. These three alleles (as well as rarer alleles) occur in virtually all human populations, with the frequency of the alleles varying from population to population. The e3 allele is commonest all populations, while the frequency of e2



and e4 varies. Numerous studies have demonstrated association between ApoE alleles and risk of various diseases or biochemical abnormalities. For example the e4 allele is associated with risk of late onset Alzheimer's disease and elevated serum cholesterol.

It has been apparent for several years that the e2, e3, e4 classification does not provide sufficient sensitivity or specificity to be used alone as a diagnostic test for assessing risk of or making a diagnosis of either dyslipidemia, heart disease or Alzheimer's disease (AD) in asymptomatic individuals. Even the use of ApoE testing as a tool in the differential diagnosis of dementia (e.g. to increase the certainty of a clinical diagnosis of Alzheimer's type dementia in a patient with early signs of dementia in whom the diagnosis of Alzheimer's is being considered) is debated. Thus, while many important associations between ApoE genotype and medically important conditions or treatment responses have been described and repeatedly confirmed, it is evident that the strength of these associations is not as great as would be desirable for a routine predictive, diagnostic or prognostic test, and in fact may not be sufficient to justify ApoE genetic testing for any non-research purpose.

The lack of sensitivity and specificity that limits the use of current ApoE genotype tests is likely attributable to two factors. First, the current ApoE test may not measure all the functional variation in the ApoE gene. For example, it does not take full account of any genetically determined variation in transcription regulation; variation in RNA processing - including splicing, polyadenylation and export to the cytoplasm; variation in mRNA translational efficiency and half life, as well as variation in protein activity including receptor binding, interaction with regulatory factors, half life, etc.. This is true particularly insofar as such variation may be determined by polymorphisms other than those that account for the e2, e3, e4 classification. Second, there may be variables besides ApoE allele status that affect the various conditions for which ApoE genotyping has been tested. Other relevant variables for neurodegenerative diseases such as AD include variation in the genes that encode protein components of AD lesions, such as tau protein or amyloid precursor protein; the proteases that produce pathological forms of these proteins, such as beta and gamma secretase and the memapsins; AD disease genes such as presenilin 1 and 2; genes involved in brain inflammatory response pathways, and other groups of genes implicated in neurodegeneration by biochemical, genetic or epidemiological evidence. Variables that may interact with ApoE genotype or haplotype to affect cholesterol and triglyceride levels and heart disease risk include the genes encoding ApoE receptors (low density lipoprotein receptor, and the low density lipoprotein receptor related protein), and genes encoding other apolipoproteins and their receptors, as well as the genes of cholesterol biosynthesis, including hydroxymethylglutaryl CoA reductase, mevalonate synthetase, mevalonate kinase, phosphomevalonate kinase, squalene synthase and other enzymes.

The present invention addresses the first limitation of current ApoE testing (failure of current ApoE tests to record all the alleles of ApoE that have distinct biochemical or clinical effects) by providing for a much more sensitive test of ApoE variation. Specifically, we describe 20 DNA polymorphisms in and around the ApoE gene (including the two polymorphisms that are traditionally studied). We also describe the commonly occurring haplotypes at the ApoE locus – that is, the sets of polymorphic nucleotides that occur together on individual chromosomes - and novel methods for determining haplotypes in clinical samples. Also described are data analysis strategies for extracting the maximum information from the ApoE haplotypes, so as to enhance their utility in clinical settings.

The ApoE haplotypes include any haplotype that can be assembled from the sequence polymorphisms described herein in Table 2, or any subset of those polymorphisms. Thus, the invention expressly includes a haplotype including either of the alternative nucleotides at any 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 of the identified polymorphic sites. The haplotypes expressly include each combination of sites with each selection of alternative nucleotide at each site included in the haplotype. The haplotypes may also include one or more additional polymorphic sites which are known in the art or which may be identified in the future. Among the haplotypes described below are a set of haplotypes that parallel the current e2, e3, e4 classification but do not involve either of the nucleotides that specify the e2, e3, e4 system.

The present invention also addresses the second potential limitation of current ApoE testing - failure to test for the interaction of ApoE genotype or haplotype with other genetic determinants of nervous system disease or cardiovascular disease risk, prognosis or response to therapy. The phenotypes for which ApoE genotyping or haplotyping have been tested are determined by multiple genes, and therefore require the simultaneous analysis of variation in two or more genetic loci. The haplotyping methods of this application facilitate such analysis by providing a basis for (i) identifying substantially all haplotypes that exist at appreciable frequency in a population or populations, (ii) clustering said haplotypes in groups of two or more haplotypes to facilitate statistical analysis, thereby increasing the power of association studies.

As used herein, "population" refers to a group of individuals that share geographic (including, but not limited to, national), ethnic or racial heritage. A population may also comprise individuals with a particular disease or condition ("disease population"). The concept of a population is useful because the occurrence and/or frequency of DNA polymorphisms and haplotypes, as well as their medical implications, often differs between populations. Therefore knowing the population to which a subject belongs may be useful in interpreting the health consequences of having specific haplotypes. A population preferably encompasses at least ten thousand, one hundred thousand, one million or more individuals, with the larger numbers being more preferable. In embodiments of this invention, the allele (haplotype) frequency,

heterozygote frequency, or homozygote frequency of a two or more alleles of a gene or genes is known in a population. In preferred embodiments of this invention, the frequency of one or more variances that may predict response to a treatment is determined in one or more populations using a diagnostic test.

5 In one aspect, the invention provides a method for determining a genotype for ApoE in an individual, comprising determining the nucleotide present at least one polymorphic site different from nucleotides 21250, and 21388 in an ApoE allele from an individual. In preferred  
embodiments, the polymorphic site is selected from the group consisting of nucleotides 16541, 16747, 16965, 17030, 17098, 17387, 17785, 17874, 17937, 18145, 18476, 19311, 20234, 21349,  
10 23524, 23707, 23759, 23805, and 37237. In certain embodiments, the method also comprises determining the nucleotide present at at least one of nucleotides 21250 and 21388. The determining is performed by a method comprising variance specific nucleic acid hybridization. The variance specific nucleic acid hybridization can be performed on an array, preferably an array composed of immobilized oligonucleotides or in situ synthesized oligonucleotides and the hybridizing species are DNA fragments. In certain embodiments, the DNA fragments are PCR amplification products. In some embodiments, the array is composed of immobilized DNA fragments and the hybridization species are oligonucleotides.

Determining the nucleotide present at a polymorphic site can be performed using a primer extension method distinguishing between nucleotides present at said at least one site, for example, as method using dideoxynucleotides to effect nucleic acid chain termination. That determining can alternatively be performed using a method involving chemical cleavage of a nucleic acid molecule including a said polymorphic site. The nucleic acid fragment masses following said chemical cleavage is preferably determined using mass spectrometry.

15 In other embodiments, determining the nucleotide present at a polymorphic site is performed using an cleavase based signal amplification method.

The nucleotide determination can also be performed using a bead-based method, preferably where the beads have a bound oligonucleotide species which is perfectly matedched or one base mismatched to the target.

Again alternatively, the determining can be performed using a FRET-based method.

20 In another aspect, the invention provides a method for determining a haplotype for ApoE in an individual, by genotyping at least two polymorphic sites in ApoE sequence on at least one allele of said individual, preferably where at least one of said polymorphic sites is different from nucleotides 21250 and 21388. As in the preceding aspect, in preferred embodiments, the  
35 polymorphic sites include at least one site selected from the group consisting of nucleotides

16541, 16747, 16965, 17030, 17098, 17387, 17785, 17874, 17937, 18145, 18476, 19311, 20234, 23524, 23707, 21349, 23759, 23805, and 37237.

In preferred embodiments, the genotyping is performed on two alleles of said individual.

In preferred embodiments, the genotyping is performed for at least

5 3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19, or 20 of the polymorphic sites.

Embodiments of the preceding two aspects can also be applied in connection with additional aspects, particularly aspects concerning ApoE described herein.

10 The invention also provides a method for classifying ApoE haplotypes for a plurality of individuals, by determining at least one ApoE haplotype for each of the plurality of individuals, determining the sequence similarity of the haplotypes (using methods for determining sequence similarity as known to those of ordinary skill in the art, and assigning the haplotypes to groups of haplotypes based on said sequence similarities. This method thus constructs groups of related ApoE haplotypes based on sequence relationship.

Further, the invention provides a method for providing an indication of the risk for an individual to develop a disease or condition, by determining a haplotype of ApoE in the individual, where the haplotype provides a measure of the risk.

In preferred embodiments of this aspect and other aspects relating to ApoE and a disease, the disease is selected from the group consisting of coronary heart disease, a non-Alzheimer's Disease neurological disease, Alzheimer's disease, stroke, brain trauma, amyotrophic lateral sclerosis, temporal lobe epilepsy, Wilson's disease, continuous ambulatory peritoneal dialysis, glycogen storage disease type Ia, and age-related macular degeneration.

25 The method (and other methods described herein relating to ApoE and disease) can also include determining a genotype or haplotype of at least one additional gene, where the haplotype of ApoE together with the genotype or haplotype of the additional gene(s) provides a measure of the risk.

30 The invention provides a method for diagnosing the presence of a disease in an individual, by determining whether the individual has an ApoE haplotype associated with the disease.

In preferred embodiments, the method also includes determining a genotype or haplotype of at least one additional gene and determining whether the individual has a combination of the

haplotype of ApoE and the genotype or haplotype of the at least one additional gene associated with the disease.

Likewise, the invention provides a method for predicting the clinical course for a patient suffering from a disease, by determining an ApoE haplotype for the individual, where at least one ApoE haplotype is associated with the clinical course of the disease.

In preferred embodiments, the clinical course comprises a treatment prognosis for a particular method of treatment, the clinical course comprises at least one clinical disease parameter selected from the group consisting of rate of disease development, time interval to death, time interval to dementia, and time interval to inability to live independently.

The invention also provides a method for selecting a subject for prophylactic treatment of a disease, by identifying a subject having an ApoE haplotype associated with an elevated risk of developing the disease, wherein said prophylactic treatment can provide a clinical benefit to a the subject.

The invention also provides a method for selecting a patient for treatment of a disease, involving determining whether the patient has an ApoE haplotype associated with favorable clinical prognosis with a particular treatment.

Similarly, the invention provides a method for selection of a treatment for a patient suffering from a disease. The method involves determining an ApoE haplotype for the patient; and identifying a treatment associated with favorable clinical prognosis for a patient having that ApoE haplotype.

As ApoE haplotype is associated with treatment selection and prognosis, the invention also provides a method of treating a patient suffering from a disease, by determining an ApoE haplotype for the patient, identifying a treatment associated with favorable clinical prognosis for a patient having that ApoE haplotype, and administering that treatment to the patient.

ApoE haplotype and genotype information also can be utilized in identifying individuals, or the individual source of a biological sample. Thus, the invention provides a method for determining whether a biological sample was from an individual, by determining the nucleotides present at a plurality of ApoE polymorphic sites in the individual and in DNA obtained from the sample, and determining whether the nucleotides present at the polymorphic sites are the same or different. The presence of the same nucleotides at respective sites is indicative that said sample

is from said individual, and the presence of different nucleotides is indicative that said sample is not from said individual. The ApoE genotype or haplotype information can also be usefully combined with similar information for polymorphic sites in other genes or other nucleic acid sequences from the individual and the sample. In preferred embodiments, the plurality of ApoE polymorphic sites comprises an ApoE haplotype.

The invention also provides a method for determining whether an ApoE haplotype is associated with a disease risk. This method involves determining ApoE haplotypes for each individual in a set of individuals, dividing the set of individuals into at least two groups based on ApoE haplotypes, and determining whether individuals having a particular ApoE haplotype or individuals in a group differ from individuals having a different ApoE haplotype or in a different group in incidence, prevalence, severity, or progression or a combination thereof, of disease. This aspect can also be combined with embodiments of other aspects described herein involving ApoE and disease, disease treatment and other such aspects.

The invention also provides a method for determining whether a combination of an ApoE haplotype and a genotype or haplotype of at least one additional gene is associated with a disease risk. The method includes determining ApoE haplotypes and genotypes or haplotypes for the at least one additional gene for each individual in a set of individuals, dividing the set of individuals into at least two groups based on the combinations of ApoE haplotypes and genotype or haplotype of said at least one additional gene, and determining whether individuals having a particular combination or individuals in a group differ from individuals having a different combination or in a different group, in incidence, prevalence, severity, or progression or a combination thereof, of said disease.

The invention further provides a method for determining whether an ApoE haplotype is associated with a pharmacologic parameter, by measuring the parameter for cells of at least one individual with said ApoE haplotype, measuring the parameter for cells of at least one individual with a different ApoE haplotype, and comparing the measures. Preferably a larger number, e.g., at least 3, 5, 10, 20, 30, 50, 100, or even more, of individuals are utilized, thereby providing additional correlation information. Correlation or other statistical measure of relatedness between haplotype and pharmacologic parameter can be used by one of ordinary skill in the art.

As used herein "polymorphism" refers to DNA sequence variation in the cellular genomes of plants or animals, preferably mammals, and more preferably humans. These sequence variations include mutations, single nucleotide changes and insertions and deletions.

“Single nucleotide polymorphism” (SNP) refers to those differences among samples of DNA in which a single nucleotide base pair has been substituted by another.

As used herein “variance” or “variants” is synonymous with polymorphism, and refers to DNA sequence variations. The terms “variant form of a gene”, “form of a gene”, or “allele” refer to one specific sequence of a gene that has at least two sequences, the specific forms differing from other forms of the same gene at at least one, and frequently more than one, variant sites within the gene. The sequences at these variant sites that differ between different alleles of the gene are variously termed “alleles”, “gene sequence variances”, “variances” or “variants”. The term “alternative form” refers to an allele that can be distinguished from other alleles by having distinct variances at least one, and frequently more than one, variant sites within the gene sequence. Other terms known in the art to be equivalent include mutation and polymorphism, although mutation is often used to refer to an allele associated with a deleterious phenotype.

As used herein “phenotype” refers to any observable or otherwise measurable physiological, morphological, biological, biochemical or clinical characteristic of an organism. The point of genetic studies is to detect consistent relationships between phenotypes and DNA sequence variation (genotypes). DNA sequence variation will seldom completely account for phenotypic variation, particularly with medical phenotypes of interest (e.g. commonly occurring diseases). Environmental factors are also frequently important.

As used herein, “genotype” refers to the genetic constitution of an organism. More specifically, “genotyping” as used herein refers to the analysis of DNA in a sample obtained from a subject to determine the DNA sequence in a specific region of the genome - e.g. at a gene that influences a disease or drug response. The term “genotyping” may refer to the determination of DNA sequence at one or more polymorphic sites.

As used herein, “haplotype” refers to the partial or complete sequence of a segment of DNA from a single chromosome. The DNA segment may include part of a gene, an entire gene, several genes, or a region devoid of genes (but which perhaps contains DNA sequence that regulates the function of nearby genes). The term “haplotype”, then, refers to a *cis* arrangement of two or more polymorphic nucleotides on a particular chromosome, e.g., in a particular gene. The haplotype preserves information about the phase of the polymorphic nucleotides – that is, which set of variances were inherited from one parent (and are therefore on one chromosome), and which from the other. A genotyping test does not provide information about phase. For example, a subject heterozygous at nucleotide 25 of a gene (both A and C are present) and also at nucleotide 100 of the same gene (both G and T are present) could have haplotypes 25A – 100G and 25C – 100T, or alternatively 25A – 100T and 25C – 100G. Only a haplotyping test can discriminate these two cases definitively. Haplotypes are generally inherited as units, except in the event of a recombination during meiosis that occurs within the DNA segment spanned by the

haplotype - a rare occurrence for any given sequence in each generation. By "haplotyping", or "determining the haplotype" as used herein is meant determining the sequence of two or more polymorphic sites on a single chromosome. Usually the sample to be haplotyped consists initially of two admixed copies of the chromosome segment to be haplotyped - i.e. DNA from a diploid subject.

As used herein "genetic testing" or "genetic screening" refers to the genotyping or haplotyping analyses performed to determine the alleles present in an individual, a population, or a subset of a population.

"Disease risk" as used herein refers to the probability that, for a specific disease (e.g. coronary heart disease) an individual who is free of evident disease at the time of testing will subsequently be affected by the disease.

"Disease diagnosis" as used herein refers to ability of a clinician to appropriately determine and identify whether the expressed symptomatology, pathology or physiology of a patient is associated with a disease, disorder, or dysfunction.

"Disease prognosis" as used herein refers to the forecast of the probable course and outcome of a disease, disorder, or dysfunction.

"Therapeutic management" as used herein refers to the treatment of disease, disorders, or or dysfunctions by various medical methods. By "disease management protocol" or "treatment protocol" is meant a means for devising a therapeutic plan for a patient using laboratory, clinical and genetic data, including the patient's diagnosis and genotype. The protocol clarifies therapeutic options and provides information about probable prognoses with different treatments. The treatment protocol may provide an estimate of the likelihood that a patient will respond positively or negatively to a therapeutic intervention. The treatment protocol may also provide guidance regarding optimal drug dose and administration, and likely timing of recovery or rehabilitation. A "disease management protocol" or "treatment protocol" may also be formulated for asymptomatic and healthy subjects in order to forecast future disease risks based on laboratory, clinical and genetic variables. In this setting the protocol specifies optimal preventive or prophylactic interventions, including use of compounds, changes in diet or behavior, or other measures. The treatment protocol may include the use of a computer program.

The term "associated with" in connection with the relationship between a genetic characteristic, e.g., a gene, allele, haplotype, or polymorphism, and a disease or condition means that there is a statistically significant level of relatedness between them based on any generally accepted statistical measure of relatedness. Those skilled in the art are familiar with selecting an appropriate statistical measure for a particular experimental situation or data set. The genetic characteristic, e.g., the gene or haplotype, may, for example, affect the incidence, prevalence, development, severity, progression, or course of the disease. For example, ApoE or a particular



allele(s) or haplotype of the gene is related to a disease if the ApoE gene is involved in the disease or condition as indicated, or if a particular sequence variance, haplotype, or allele is so involved.

As used herein, a "gene" is a sequence of DNA present in a cell that directs the expression of a "biologically active" molecule or "gene product", most commonly by transcription to produce RNA and translation to produce protein. Such a gene may also be manipulated by many different molecular biology techniques, and thus, for example, can be isolated or purified or otherwise separated from its natural environment. The "gene product" is most commonly a RNA molecule or protein or a RNA or protein that is subsequently modified by reacting with, or combining with, other constituents of the cell. Such modifications may include, without limitation, modification of proteins to form glycoproteins, lipoproteins, and phosphoproteins, or other modifications known in the art. RNA may be modified without limitation by polyadenylation, splicing, capping or export from the nucleus or by covalent or noncovalent interactions with proteins. The term "gene product" refers to any product directly resulting from transcription of a gene. In particular this includes partial, precursor, and mature transcription products (i.e., pre-mRNA and mRNA), and translation products with or without further processing including, without limitation, lipidation, phosphorylation, glycosylation, or combinations of such processing

As used herein the term "hybridization", when used with respect to DNA fragments or polynucleotides encompasses methods including both natural polynucleotides, non-natural polynucleotides or a combination of both. Natural polynucleotides are those that are polymers of the four natural deoxynucleotides (deoxyadenosine triphosphate [dA], deoxycytosine triphosphate [dC], deoxyguanine triphosphate [dG] or deoxythymidine triphosphate [dT], usually designated simply thymidine triphosphate [T]) or polymers of the four natural ribonucleotides (adenosine triphosphate [A], cytosine triphosphate [C], guanine triphosphate [G] or uridine triphosphate [U]). Non-natural polynucleotides are made up in part or entirely of nucleotides that are not natural nucleotides; that is, they have one or more modifications. Also included among non-natural polynucleotides are molecules related to nucleic acids, such as peptide nucleic acid [PNA]). Non-natural polynucleotides may be polymers of non-natural nucleotides, polymers of natural and non-natural nucleotides (in which there is at least one non-natural nucleotide), or otherwise modified polynucleotides. Non-natural polynucleotides may be useful because their hybridization properties differ from those of natural polynucleotides. As used herein the term "complementary", when used in respect to DNA fragments, refers to the base pairing rules established by Watson and Crick: A pairs with T or U; G pairs with C. Complementary DNA fragments have sequences that, when aligned in antiparallel orientation, conform to the Watson-Crick base pairing rules at all positions or at all positions except one. As

used herein, complementary DNA fragments may be natural polynucleotides, non-natural polynucleotides, or a mixture of natural and non-natural polynucleotides.

As used herein "amplify" when used with respect to DNA refers to a family of methods for increasing the number of copies of a starting DNA fragment. Amplification of DNA is often performed to simplify subsequent determination of DNA sequence, including genotyping or haplotyping. Amplification methods include the polymerase chain reaction (PCR), the ligase chain reaction (LCR) and methods using Q beta replicase, as well as transcription-based amplification systems such as the isothermal amplification procedure known as self-sustained sequence replication (3SR, developed by T.R. Gingeras and colleagues), strand displacement amplification (SDA, developed by G.T. Walker and colleagues) and the rolling circle amplification method (developed by P. Lizardi and D. Ward).

By "comprising" is meant including, but not limited to, whatever follows the word "comprising". Thus, use of the term "comprising" indicates that the listed elements are required or mandatory, but that other elements are optional and may or may not be present. By "consisting of" is meant including, and limited to, whatever follows the phrase "consisting of". Thus, the phrase "consisting of" indicates that the listed elements are required or mandatory, and that no other elements may be present. By "consisting essentially of" is meant including any elements listed after the phrase, and limited to other elements that do not interfere with or contribute to the activity or action specified in the disclosure for the listed elements. Thus, the phrase "consisting essentially of" indicates that the listed elements are required or mandatory, but that other elements are optional and may or may not be present depending upon whether or not they affect the activity or action of the listed elements.

Other features and advantages of the invention will be apparent from the following description of the preferred embodiments thereof, and from the claims.

## DETAILED DESCRIPTION OF THE INVENTION

### Brief description of the Figures and Tables

**Table 1** The table lists the masses of the normal nucleotides and BrdU and the mass differences between each of the possible pairs of nucleotides.

**Table 2** Twenty polymorphic sites in the ApoE gene. The ApoE genomic sequence is taken from GenBank accession AB012576. The gene is composed of four exons and three introns. The transcription start site (beginning of first exon) is at nucleotide (nt) 18,371 of GenBank accession AB012576, while the end of the transcribed region (end of the 3' untranslated region, less polyA tract) is at nt 21958. The twenty polymorphic sites are depicted as shaded nucleotides in the Table, and are as follows (nucleotide position and possible nucleotides): 16541 (T/G); 16747 (T/G); 16965 (T/C); 17030 (G/C); 17098 (A/G); 17387 (T/C); 17785 (G/A); 17874 (T/A); 17937 (C/T); 18145 (G/T); 18476 (G/C); 19311 (A/G); 20334 (A/G); 21250 (C/T); 21349 (T/C); 21388 (T/C); 23524 (A/G); 23707 (A/C); 23759 (C/T); 23805 (G/C); and 37237 (G/A). The bold sequence listing indicates the transcribed sequence of the ApoE gene; the grey shaded region indicates the ApoE gene enhancer element; the underlined sequence depicts the coding region of the ApoE gene. Where polymorphisms result in a change of the amino acid sequence, the amino acid alteration is indicated, for example at nucleotide position 20334 the A/T polymorphism results in a alanine/threonine respectively at amino acid position 18 of the ApoE gene product. As described in the Detailed Description below, the polymorphisms at positions GenBank nucleotide number 17874, 17937, 18145, 18476, 21250, and 21388 have been previously described.

**Table 3** This table provides experimentally derived ApoE haplotypes. The haplotypes encompass nine polymorphic sites within the ApoE gene (GenBank accession number AB012576). The Table has nine columns with haplotype data at nine specific sites within the ApoE gene. The column listed as "WWP #" refers to a Coriell number which refers to the catalogued number of an established human cell line. The "VGNX\_Symbol" row provides an internal identifier for the gene; the "VGNX database" row identifies the base pair number of the ApoE cDNA; and the "GenBank" row identifies the GenBank base pair number of the sequence for the ApoE gene. The abbreviations are as follows: A=adenine nucleotide, C=cytosine nucleotide, G=guanosine nucleotide, and T=thymidine nucleotide. The abbreviated nucleotides in brackets indicate that either nucleotide may be present in the sample. Thus for example, under column GEN-CBX and **WWP#1**, the genotype identified at the GenBank position 17874 is an "A"; whereas under Column GEN-CBX at the GenBank position 18476 the genotype under the **WWP#1** is either a "T" or a "G".

**Table 4** This table provides the sequence of ApoE haplotypes comprising up to 20 polymorphic sites. There are 42 ApoE haplotypes listed in the Table. The top row of the table provides the location of the polymorphic nucleotides in the ApoE gene (see Table 2). The numbers (16541, 16747, and so forth) correspond to the numbering in GenBank accession AB012576\_1, which provides the sequence of a cosmid clone that contains the entire ApoE gene and flanking DNA. Each column shows the sequence of the ApoE gene at the position indicated at the top of the column. Abbreviations are as follows: A=adenine nucleotide, C=cytosine nucleotide, G=guanosine nucleotide, and T=thymidine nucleotide. Each row provides the sequence of an individual phenotype.

**Table 5** This table provides the sequence of haplotypes at the the ApoE gene determined by 5 polymorphic sites. These haplotypes allow classification of ApoE alleles into the e2, e3 and e4 groups without recourse to the polymorphic sites conventionally used to determine e2, e3, e4 status. In this table the haplotypes are specified by SNPs at positions 16747, 17030, 17785, 19311, and 23707, listed as column headings. The GENOTYPE column provides the classic ApoE genotype/phenotype (e2, e3 and e4) corresponding to the haplotype indicated in each row.

**Figure 1** Depiction of a primer designed to incorporate restriction enzyme recognition sites for the specific restriction enzymes Fok I and Fsp I. The primer (primer R sequence) has altered bases from the desired amplified region of the target DNA. The polymorphic nucleotide is included in the target DNA region and is as indicated by the arrow. After PCR amplification, the incorporated altered base pairs of the primer is thereby incorporates a FokI and FspI restriction sites in the amplicon. The presence of the FokI and FspI sites can subsequently be digested in the presence of the FokI and FspI restriction enzymes under optimal conditions for digestion by both enzymes. The resultant fragments, an 8 mer and a 12 mer, after enzyme digestion are as depicted. In this figure, the polymorphism (A, in *italic*) is contained with the 12 mer fragment.

**Figure 2** This figure depicts the utility of Fok I, a type IIS restriction enzyme, which cleaves DNA outside the recognition sequence at a distance of 9 bases 3' to the recognition site on one strand and 13 bases away from the recognition site on the opposite strand, leaving a four base overhang (protruding 5' end). As shown in this figure, by designing the primer so that the Fok I recognition site is located within 12 bases or less of the 3' end of the primer one can assure that the Fok I cleavage will cleave outside the primer sequence. Further shown is the utility of FspI, a restriction enzyme that after digestion leaves blunt ends. The FspI recognition site, TGCGCA, after digestion results in fragments as shown.

**Figure 3** In this figure, the utility of the Fsp I/Fok I pair of enzymes for the present invention is shown. The FspI recognition site overlaps that of Fok I, allowing the two sites to be partially combined. Thus, including the combined FspI/FokI sequence in the primer, reduces the number of bases that are to be introduced into the modified primer, making the primer design simpler and more likely to function in the subsequent amplification reaction.

**Figure 4** In this figure, an alternative method of primer design in the present invention involves the use of a primer with an internal loop. The primer is designed (primer R1) such that one of the bases corresponding to the native sequence is removed and replaced with a loop. In this case the G/C indicated by the arrow below the target sequence is replaced with the recognition sequence for Fok I and Fsp I. Upon hybridization to the DNA template, the primer will form a loop structure. This loop will be incorporated into the amplicon during the amplification process, thereby introducing the Fok I and Fsp I restriction sites (indicated by the box). The resultant amplicon is incubated with Fok I and Fsp I under optimal digestion conditions producing an 8-mer and a 12-mer fragment. As in Figure 1, the 12-mer contains the polymorphic base (A in *italic*) and can be analyzed by mass spectrometry to identify the base at the polymorphic site.

**Figure 5** Alternative restriction enzyme recognition site incorporation into amplified regions of target DNA is shown. As is depicted in figures 1-4 for the enzyme pair FspI/FokI; in this figure, PvuII/FokI restriction enzymatic sites can be incorporated in the same manner as previously described for Figures 1-4. A primer is designed such that a BsgI/PvuII sites form a hair-pin loop when the primer is hybridized to the target DNA sequence. After amplification by PCR, the resultant amplicon will have the PvuII/FokI sites incorporated in the resultant amplicon (as indicated by the boxed sequence). After digestion under conditions optimal for PvuII and BsgI, the resultant fragments, an 14 mer and a 16 mer, are sufficient for mass spectrometric analysis and the polymorphic site is contained in the 16mer (A, in *italic*).

**Figure 6** Shown in this figure is an alternative restriction enzyme pair for the preparation of fragments containing the polymorphic site for mass spectrometric analysis. As shown in the figure, the primer has a PvuII/FokI restriction enzyme recognition sites that form a hair-pin loop when hybridized to the target DNA sequence. After amplification by PCR, the resultant amplicon will have the PvuII/FokI sites incorporated in the resultant amplicon (as indicated by the boxed sequence). After digestion under conditions optimal for PvuII and FokI, the resultant

fragments, an 16 mer and a 20 mer, are sufficient for mass spectrometric analysis and the polymorphic site is contained in the 20mer (A, in *italic*).

**Figure 7** In this figure, a modification of the method depicted in Figure 4 is shown. As in Figure 4, a DNA segment containing a polymorphism is amplified using two primers. One primer is designed with an inserted DNA segment, not complementary to template DNA, that forms a hair-pin loop when hybridized to template DNA. Insertion of the non-complementary DNA segment results in incorporation of overlapping FokI and FspI restriction enzyme sites after PCR amplification ( as shown in the boxed sequence). Following PCR amplification reaction, the reaction is subjected to a clean up procedure to remove unincorporated primers, nucleotides and buffer constituents. The PCR product is then digested with the FokI restriction enzyme which generates a 5' overhang that extends from the 3' end of the primer to beyond the polymorphic nucleotide. The 3' recessed end can then be filled in with exogenously added nucleotides in which the normal nucleotide corresponding to one of the possible nucleotide bases at the polymorphic site is a mass modified nucleotide ( $T^{mod}$ ). These fragments are sufficient for mass spectrometric analysis of the modified polymorphic nucleotide.

**Figure 8** Shown in this figure is the incorporation of a single restriction enzyme recognition site in the amplicon for subsequent digestion and mass spectrometric analysis of the prepared fragments. Shown in this figure is incorporation of BcgI, an restriction enzyme that is capable of making two double strand cuts, one on the 5' side and one on the 3' side of their recognition site. The recognition site for BcgI is 12/10(N)CGA(N)<sub>6</sub>TGC(N)12/10, which after digestion results in fragments sufficient for mass spectrometric analysis and identification of the polymorphic base with the fragment.

**Figure 9** Shown in this figure is an example of the utility in the present invention of including a restriction enzyme recognition site for which the restriction enzyme creates a nick in the DNA amplicon instead of causing a double strand break. As shown in this figure, a primer R is designed to incorporate a N.BstNB I recognition site (GAGTCNNNN<sup>^</sup>NN) in addition to a FokI restriction site. As in previous figures, the primer forms a hair-pin loop structure when hybridized to the target DNA region, however, the PCR amplicon has the incorporated restriction site sequences. Digestion with FokI and N.BstNB I results in a 10 mer fragment that contains the polymorphic base (*T* in *italic*). Such a fragment is sufficient for analysis using a mass spectrometer.

**Figure 10** Shown in this figure is a similar strategy to the nicking enzyme scheme of Figure 9, above. In this method, one restriction enzyme and a primer which contains a ribonucleotide substitution for one of the deoxyribonucleotides. As shown the primer is designed to contain a FokI recognition site which upon hybridization with the target DNA sequence forms a hair-in

5 loop. The primer also has a ribonucleoside (rG) substitution which will additionally be incorporated into the amplicon. The ribonucleoside substitution is base-labile and will cause a break in the backbone of the DNA at that site under basic conditions. Shown in this scheme, the amplicon is incubated with the restriction enzyme (Fok I) causing a double-strand break. The amplicon is then incubated in the presence of base causing a break between the ribonucleotide G

10 and the 3' deoxyribonucleotide T, releasing a 7 base fragment which can easily analyzed by mass spectrometry.

**Figure 11** The diagram illustrates the major approaches to haplotyping within the allele separation group of allele enrichment methods, described in section A of the specification. As shown, methods can be broadly categorized as (1) those directed to single stranded DNA and (2) those directed to double stranded DNA. It is possible to capture DNA fragments in an allele specific manner by affinity to proteins or nucleic acids that discriminate single base differences. Different types of protein and nucleic acid affinity reagents are shown in the boxes. The protein or nucleic acid that sticks to one allele can subsequently be selected from the nucleic acid mixture by methods known in the art such as streptavidin or antibody coated beads. A third, non-affinity based method for separating alleles involves restriction endonuclease cleavage at a polymorphic site (such that fragments of significantly different size are produced from the two alleles), and subsequent size fractionation of the cleaved products using electrophoresis or centrifugation. Genotyping the isolated fragments corresponding to the two alleles will provide

25 haplotypes.

**Figure 12** This diagram depicts the various methods one skilled in the art could employ for haplotyping based on allele-specific amplification. After cleavage of one allele the other allele may be selectively amplified, or separated by a size selection procedure, or the cleaved allele may

30 be removed by an allele selective degradation procedure.

**Figure 13** This diagram depicts the categorization of the various methods one could employ haplotyping strategies based upon allele specific restriction. In these methods one allele is preferentially amplified from a mixture of two alleles by the design of a primer or primers that

35 exploit sequence differences at polymorphic sites.

**Figure 14** Hair pin loop primers. In this figure the primers used for PCR amplification is shown. In allele 1, the polymorphic site is a T (*italic*) and incorporation of the ATCTGGA 5' portion of the primer occurs after at least one round of amplification. In allele 2, the polymorphic site is a C (*italic*) and incorporation of the ATCTGGA 5' portion of the primer occurs at least after one round of amplification.

**Figure 15** Hair pin loop primers. In this figure the primers used for PCR amplification is shown. In allele 1, the polymorphic site is a C (*italic*) and incorporation of the ATCCGGA 5' portion of the primer occurs after at least one round of amplification. In allele 2, the polymorphic site is a T (*italic*) and incorporation of the ATCCGGA 5' portion of the primer occurs at least after one round of amplification.

**Figure 16** Hair pin loop primers. In this figure, the minus strand of allele 1 generated by the PCR amplification step shown in Figure 14 depicts the inability of the 5' primer to hybridize and effectively prevents the amplification of allele 1, using the T primer. Alternatively, the minus strand of allele 2 is incapable of forming a hairpin loop due to the mismatch. Thus, hairpin loop formation and prevention of PCR amplification does not occur, and amplification this allele 2 strand will occur using the T primer.

**Figure 17** Hair pin loop primers. In this figure, the minus strand of allele 2 generated by the PCR amplification step shown in Figure 15 depicts the inability of the 5' primer to hybridize and effectively prevents the amplification of allele 2, using the C primer. Alternatively, the minus strand of allele 1 is incapable of forming a hairpin loop due to the mismatch. Thus, hairpin loop formation and prevention of PCR amplification does not occur, and amplification the allele 1 strand will occur using the C primer.

**Figure 18** Exonuclease based methods for the determination of a haplotype. In the DNA segment to be haplotyped, one identified site of polymorphism is a RFLP, so that on one allele the restriction enzyme, as shown as BamHI is able to digest the alleles and generate different length fragments.

**Figure 19** Exonuclease based method for the determination of a haplotype. Using the fragments as shown and described in figure 16, the ends of the DNA fragments are protected from exonuclease digestion. The protected fragments are then digested with a second restriction enzyme for whose recognition site is located in one of the fragments, but not the other, due to the



overhang of the RFLP, as shown, a *NheI* site. Restriction digestion of the fragments with *NheI* will effectively shorten the *BamHI* fragment but additionally remove the protection from the exonuclease digestion.

5 **Figure 20** Endonuclease based method for the determination of a haplotype. Using the fragments generated as shown in figure 17, these fragments are then incubated in the presence of an exonuclease. As shown the exonuclease will digest one of the fragments but the protected fragments will remain undigested.

10 **Figure 21** Primer mediated inhibition of allele-specific PCR amplification. Primers with the above characteristics were designed for haplotyping of the dihydropyrimidine dehydrogenase (DPD) gene. The DPD gene has two sites of variance in the coding region at base 186 (T:C) and 597 (A:G) which result in amino acid changes of Cys:Arg and Met:Val, respectively as shown in the box of Figure 27. The second site at base 597 is a restriction fragment length polymorphism (RFLP) which cleaves with the enzyme *BsrD I* if the A allele is present. The expected fragments are as shown in the figure.

**Figure 22** Allele specific primers for the DPD gene. In A., three primers were designed which contain at least two different regions. The 3' portion of the primer corresponds to the template DNA to be amplified. For the DPDASCF and the DPDASTF primers additional nucleotides were added to the 5' end of the primer which are complementary to the region in the sequence which contains the nucleotide variance. The DPDNSF primer contains only the DPD complementary sequence and will not result in allele specific amplification. In B., the DPD gene sequence containing the site of polymorphism is shown.

**Figure 23** PCR amplification of the DPD gene using the DPDNSF primer. Shown is the hybridization of the DPDNSF primers to the template containing the T or C allele. Below, the expected products for the DPD gene region using the DPDNSF primer for the T or C allele as shown.

**Figure 24** PCR amplification of the DPD gene using the DPDASTF primer. Shown is the hybridization of the DPDASTF primers to the template containing the T or C allele. Below, the expected products for the DPD gene region using the DPDASTF primer for the T or C allele as shown.

**Figure 25** PCR amplification of the DPD gene using the DPDASCF primer. Shown is the hybridization of the DPDASCF primers to the template containing the T or C allele. Below, the expected products for the DPD gene region using the DPDASCF primer for the T or C allele as shown.

**Figure 26** Stable hairpin loop structures formed with the reverse strand of the PCR product made using the DPDNSF primer using the computer program Oligo4. Only the reverse strand is shown because this would be the strand to which the DPDNSF primer would hybridize on subsequent rounds of amplification. The hairpin loops are either not stable or have a low melting temperature..

**Figure 27** Stable hairpin loop structures formed with the reverse strand of the PCR product made using the DPDASCF primer using the computer program Oligo4. As in Figure 32, only the reverse strand is shown.

**Figure 28** Stable hairpin loop structures formed with the reverse strand of the PCR product made using the DPDASTF primer using the computer program Oligo4. As in Figure 32, only the reverse strand is shown.

**Figure 29** The primer hybridization and amplification events when further amplification using the DPDNSF primer is attempted on the generated PCR fragments. The DPDNSF primer is able to effectively compete with the hairpin structures formed with both the T and C allele of the DPD gene and thus amplification of both alleles proceeds efficiently.

**Figure 30** The primer hybridization and amplification events when further amplification using the DPDASCF primer is attempted on the generated PCR fragments. The DPDASCF primer is able to compete for hybridization with the hairpin loop formed with the C allele because its melting temperature is higher than the hairpin loop's ( $60^{\circ}\text{C}$  compared to  $42^{\circ}\text{C}$ ). The hairpin loop formed on the T allele however, has a higher melting temperature than the primer and thus effectively competes with the primer for hybridization. The hairpin loop inhibits PCR amplification of the T allele which results in allele specific amplification of the C allele.

**Figure 31** The primer hybridization and amplification events when further amplification using the DPDASTF primer is attempted on the generated PCR fragments. The hairpin loop structure has a higher melting temperature than the primer for the C allele and a lower melting temperature

than the primer for the T allele. This causes inhibition of primer hybridization and elongation on the C allele and results in allele specific amplification of the T allele.

**Figure 32** The ability to use the hair-pin loop formation for haplotyping the DPD gene is diagrammed. Using a cDNA sample whose haplotype is known to be : Allele 1 – T<sup>186</sup>:A<sup>597</sup>, Allele 2 – C<sup>186</sup>:G<sup>597</sup>. The size of the fragments generated by a BsrD I from a 597 bp sequence generated by amplification with the primers DPDNSF, DPDASTF, and DPDASCF, depend on whether the base at site 597 is an A or a G. Restriction digestion by BsrD I is indicative of the A base being at site 597. If a fragment has the A base at 597, three fragments will be generated of lengths 138, 164 and 267 bp. If the G base is at site 597 only two fragments will be generated of lengths 164 and 405 bp. If a sample is heterozygous for A and G at site 597, generation of all four bands of 138, 164 (2x), 267 and 405 bp will occur. The expected fragments generated by BsrD I restriction for each of the primers is indicated in the box.

**Figure 33** Agarose gel electrophoresis of the fragments generated by amplification of each of the primers for the DPD gene in a cDNA sample heterozygous at both sites 186 and 597 followed by BsrD I restriction. The DPDNSF lane shows the restriction fragment pattern for the selected cDNA using the DPDNSF primer indicating that this sample is indeed heterozygous at site 597. However, using the same cDNA sample and the primer DPDASTF (DPDASTF lane), the restriction pattern correlates to the pattern representative of a sample which is homozygous for A at site 597. Because the DPDASTF primer allows amplification of only the T allele, the haplotype for that in the sample must be T<sup>186</sup>:A<sup>597</sup>. The restriction digest pattern using the primer DPDASCF (DPDASCF lane) correlates with the expected pattern for there being G at site 597. Amplification of the cDNA sample with the primer DPDASCF results in amplification of only the C allele in the sample. Thus the haplotype for this allele must be C<sup>186</sup>:G<sup>597</sup>.

**Figure 34** Genotyping of the variance at genomic site 21250 in the ApoE gene. At this genomic site a T:C variance in the DNA results in a cysteine to arginine amino acid change in amino acid position 176 in the ApoE protein. Two primers were designed to both amplify the target region of the ApoE gene and to introduce two restriction enzyme sites (Fok I, Fsp I) into the amplicon adjacent to the site of variance. This figure depicts the sequence of the primers and the target DNA. The Apo21250-LFR primer is the loop primer which contains the restriction enzyme recognition sites and the ApoE21250-LR primer is the reverse primer used in the PCR amplification process. The polymorphic nucleotide is shown in italics.

**Figure 35** The sequence of the amplicon for both the T allele and the C allele of the ApoE gene following amplification is shown. The polymorphic site is shown as an italic T or italic C.

## 1. Genotyping Methods

This application concerns methods for determining the sequence of a DNA sample at a polymorphic site, often referred to as genotyping. Many genotyping methods are known in the art, however the method described in this application has the advantages of being robust, highly accurate, and inexpensive to set up and perform. For these reasons the novel methods described herein are preferable to most currently available methods.

The disadvantages of existing genotyping methods include: unproven or inadequate accuracy (particularly for medical research, where very high accuracy is required); high set up costs (which are unacceptable when relatively small numbers of subjects are being studied); technical difficulty in performing the test or interpreting the results; and lack of automatability.

The present invention describes a genotyping method based on mass spectrometric analysis of small DNA fragment(s) (preferably <25 bases) containing a polymorphic base. The small size of the DNA fragments generated allows them to be efficiently analyzed via mass spectrometry to determine the identity of the nucleotide at the polymorphic site. The generation of appropriate DNA fragments preferably falls in the range between 9,000 Daltons (30-mer) and about 900 Daltons (3-mer), or between 900 and 7500 Daltons (25-mer), or between 900 and 6000 Daltons (20-mer), or between 900 and 4500 Daltons (15-mer). However, as mass spectrometry technology progresses it will become possible to genotype DNA fragments outside this currently recommended range, so greater ranges are also included in preferred embodiments, e.g., 900 to 9600 Daltons (32-mer), or 900 to 10500 Daltons (35-mer), or 900 to 12000 Daltons (40-mer). Thus the methods described herein are tailored to the capabilities of presently available commercial mass spectrometers, however, one skilled in the art will recognize that these methods can be adapted with ease to improvements in mass spectrometry equipment, including, for example, MALDI instruments with improved desorption, delayed extraction or detection devices.

The invention entails use of a single modified primer in a primer extension or amplification reaction. The modified primer is designed so as to introduce at least two restriction endonuclease recognition sites into the sequence of the primer extension product, which is preferably an amplicon in an amplification reaction. The restriction endonuclease recognition sites are designed such that they surround and/or span the polymorphic base to be genotyped and will liberate a small DNA fragment(s) containing the polymorphic base upon cleavage. If the natural sequence adjacent to the polymorphic site (either on the 5' side or the 3' side) already contains a restriction endonuclease recognition site then it may be possible to design the modified primer so that one of the two restriction cleavage sites is not engineered into the primer

(see below), but rather occurs naturally in the amplicon. In this event only one restriction site has to be engineered into the primer.

One embodiment of the invention involves the introduction of two restriction enzyme sites into the sequence of an amplicon in the vicinity of a polymorphic site during amplification.

5 The two restriction enzyme sites are selected so that when the amplicon is incubated with the corresponding restriction enzymes, two small DNA fragments are generated, at least one of which contains the polymorphic nucleotide. The restriction enzyme sites are introduced during the amplification process by designing a primer that contains recognition sites for two restriction endonucleases. Two different methods for designing such primers are described below, but any  
10 strategy in which at least two cleavable sites are introduced into an amplicon using a single primer would be effective for this method.

One method involves the selected alteration of bases in the primer (relative to what they would be if the primer were to base pair perfectly with the natural sequence) so as to introduce restriction enzyme sites. An example of such a primer, incorporating recognition sites for the restriction enzymes Fok I and Fsp I, is shown in Figure 1. The recognition sites and cleavage sites for Fok I and Fsp I are depicted in Figure 2. Fok I is a type IIS restriction enzyme which cleaves DNA outside the recognition sequence - at a distance of 9 bases 3' to the recognition site on one strand and 13 bases away from the recognition site on the opposite strand, leaving a four base overhang (protruding 5' end) (Figure 2). By designing the primer so that the Fok I  
15 recognition site is located within 12 bases or less of the 3' end of the primer one can assure that the Fok I cleavage will cleave outside the primer sequence and incorporate the polymorphic nucleotide for analysis. Fsp I is a useful enzyme to pair with Fok I because its recognition site overlaps that of Fok I, allowing the two sites to be partially combined (Figure 3). This reduces the number of bases that are to be introduced into the modified primer, making the primer design  
20 simpler and more likely to work for amplification.

A primer is designed (primer R) in which some of the bases are changed from the target sequence. The bases that are changed are indicated by arrows above primer R. This primer along with a second (normal) amplification primer designed in the reverse direction are used to amplify the target sequence. The polymorphic base (T in the forward direction, A in the reverse  
25 direction) is indicated in italics and by an arrow below the target sequence. During the amplification, the two restriction enzyme sites are incorporated into the sequence of the amplicon. The incorporated Fok I/Fsp I site is surrounded by the box in Figure 1. When the amplicon is incubated with Fok I and Fsp I, cleavage occurs at the both sites releasing an 8-mer fragment and a 12-mer fragment. The 12-mer fragment contains the polymorphic base (A).

30 These fragments are then analyzed by the mass spectrometer to determine the base identity at the polymorphic site in the 12-mer.

The second method of primer design involves the use of a primer with an internal loop. The primer is designed (primer R1, Figure 4) such that one of the bases corresponding to the native sequence is removed and replaced with a loop. In this case the G/C indicated by the arrow below the target sequence (Figure 4) is replaced with the recognition sequence for Fok I and Fsp I. Upon hybridization to the DNA template, the primer will form a loop structure. This loop will be incorporated into the amplicon during the amplification process, thereby introducing the Fok I and Fsp I restriction sites (indicated by the box in Figure 4). When the amplicon is incubated with Fok I and Fsp I, cleavage will occur releasing an 8-mer and a 12-mer. As in the example in Figure 1, the 12-mer contains the polymorphic base and can be analyzed by mass spectrometry to identify the base at the polymorphic site.

Both strategies result in an amplicon which can be cleaved with Fok I and Fsp I to liberate small DNA fragments in which the polymorphic nucleotide is contained in one of the fragments. The loop strategy (Figure 4) is the preferred method because primer design is easier and more flexible.

There are other possible restriction enzyme combinations that also meet the requirements for the generation of appropriate DNA fragments for genotyping by mass spectrometry. Two other examples are outlined in Figures 5 and 6. The only requirements for primer design are that the restriction enzyme site(s) will generate a fragment(s) that is small enough to be easily analyzed by a mass spectrometer, and contain the polymorphic site. It is also a requirement that the introduction of the restriction enzyme site(s) into the primer does not eliminate the ability of the primer to generate an amplicon for the correct region of the target DNA. It does not matter whether the cleavage site for both enzymes generates a staggered 5' overhang, 3' overhang, or a blunt end.

In another embodiment it may be desirable to generate a cleavage product in which there is a 5' overhang such as the case with the Fok I and Fsp I example shown in Figure 4. Following an amplification reaction (in which the Fok I and Fsp I sites have been incorporated into the amplicon - see sequence in box Figure 7), remaining nucleotides are removed using any of a variety of methods known in the art, such as spinning through a size exclusion column such as Sephadex G50 or by incubating with an alkaline phosphatase, e.g., shrimp alkaline phosphatase. The amplicon is then cleaved with the restriction enzyme (Fok I) which generates the 5' overhang including the polymorphic base. This recessed end can then be filled in with nucleotides in which the normal nucleotide corresponding to one of the possible nucleotide bases at the polymorphic site is a mass modified nucleotide ( $T^{\text{mod}}$  in Figure 7). The mass modified nucleotide has a mass that is different from the normal nucleotides in a way that increases the difference in mass normally seen between normal nucleotides. An example of such a nucleotide is bromo-deoxyuridine (BrdU) which is 64.8 Daltons higher in mass than dTTP. Table 1 lists the

masses of the normal nucleotides and BrdU and the mass differences between each of the possible pairs of nucleotides. As is evident from the table, mass modified nucleotides allow a greater separation in mass between the fragments, making analysis, especially in an automated mode, easier. After fill-in of the recessed ends of the fragment, digestion with FspI would then allow for the generation of a fragment amenable for mass spectrometric analysis and identification of the polymorphism of interest. An advantage of this method, from the others in the present invention, allows for one to use either mass spectrometry or readily available electrophoretic detection methods. In a preferred embodiment, after digestion with FokI, and overhang fill-in reaction that includes a modified nucleotide representing one of the suspected polymorphic bases, the fragments analyzed by electrophoretic mobility would migrate differently due to the incorporation of the mass modified nucleotide. Thus, one could identify suspected polymorphic sites in the prepared unlabeled fragments by either mass spectrometric or electrophoretic methods.

Alternatively, in the above described method employing a mass modified nucleotide recessed end fill-in, a labeled primer (radioactive or fluorescent label) during the PCR reaction would result in a detectable signal if the samples were then subjected to electrophoretic separation. In this case, a target DNA sample is amplified using a similar scheme to the one described above; a 5' labeled primer with a FokI restriction site is allowed to hybridize to the target DNA forming a hair-pin loop, and subsequent amplification incorporates the FokI site into the amplicon. The resultant amplicon is subjected to digestion with FokI to separate the sequence 3' of the site of polymorphism and the residual nucleotides from the PCR reaction are removed as described above. The overhang sequence then is filled in with a polymerase in the presence of natural nucleotides with one of the nucleotides of the polymorphic site being a dideoxynucleotide, or chain terminating nucleotide. Thus, differential fill-in of the overhang will be dependent on the presence or absence of the polymorphism and thus incorporation of a dideoxy terminating nucleotide. In preferred embodiments, the primer is not labeled but the dideoxy chain terminating nucleotide representing one of the suspected polymorphic bases is labeled to be able to detect those fragments. In further preferred embodiments, each polymorphic base dideoxynucleotide is labeled with uniquely detectable labels and the identification of the polymorphic site is based upon presence of one signal and absence of another in the cases of homozygotes or the presence of both signals in the cases of heterozygotes.

It may also only be necessary to incorporate one restriction enzyme site into the amplicon via the primer. This can be done if the enzyme utilized is capable of making two double strand cuts, one on the 5' side and one on the 3' side of the recognition site. An example of such an enzyme is Bcg I which has a recognition site of 12/10(N)CGA(N)<sub>6</sub>TGC(N)12/10 (Figure 8). The arrows designate the sites of cleavage on both strands. This particular enzyme would generate

fragments greater than the current optimal length for mass spec analysis, so similar enzymes that are capable of cleaving in a similar fashion but which would generate smaller fragments are more desirable. Also, as mass spectrometry techniques and instrumentation for DNA analysis progress, it may be possible to reliably analyze DNA fragments of this length or greater obtaining the sensitivity and the resolution necessary to see single base differences in fragments of this length.

Restriction enzymes can also be used which only nick the DNA instead of causing a double strand break. One such enzyme is N.BstNB I whose recognition site is GAGTCNNNN<sup>^</sup>NN. The fragments generated by this scheme are outlined in Figure 9. This strategy would generate only one small fragment (10-mer in this case) instead of two which may make analysis less complicated, especially in an automated mode.

A similar strategy to the nicking enzyme above can be accomplished using one restriction enzyme and a primer which contains a modification allowing the primer to be cleaved. An example of such a scheme outlined in Figure 10 is where one of the deoxyribonucleosides in the primer is substituted with a ribonucleoside (rG). The ribonucleoside is base-labile and will cause a break in the backbone of the DNA at that site. In this example, the amplicon is incubated with the restriction enzyme (Fok I) causing a double-strand break. The amplicon is then incubated in the presence of base causing a break between the ribonucleotide G and the 3' deoxyribonucleotide T, releasing a 7 base fragment which can easily be analyzed by mass spectrometry.

## 2. Haplotyping Methods

### *Background*

In mammals, as in many other organisms, there are two copies (alleles) of each gene in every cell (except some genes which map to the sex chromosomes - X and Y in man). One allele is inherited from each parent. The purpose of the haplotyping methods described in this application is to determine the sequence of the two alleles in a given subject. In general the two alleles in any organism are substantially similar in sequence, with polymorphic sites occurring less than every 100 nucleotides, and in some cases in less than every 1,000 nucleotides.

Determination of the sequence of the non-variant nucleotide positions is not relevant to haplotyping. Thus, the problem of haplotyping comes down to determining the nucleotide sequence on each of the two alleles at the polymorphic sites: For a subject that is heterozygous at two sites, where polymorphic site #1 is A or C, and polymorphic site #2 is G or T, we wish to know if the alleles are A – G and C – T, or if they are A – T and C – G. When DNA is extracted from a diploid organism the two alleles are mixed together in the same test tube at a 1:1 ratio. Thus DNA analysis procedures performed on total genomic DNA, such as DNA sequencing or



standard genotyping procedures which query the status of polymorphic sites one at a time, do not provide information required to determine haplotypes from DNA samples that are heterozygous at two or more sites.

The determination of haplotypes is particularly useful for genetic analysis when the DNA segment being haplotyped consists of polymorphisms that are in some degree of linkage disequilibrium with each other – that is, they do not assort randomly in the population being studied. In general, linkage disequilibrium breaks down with increasing physical distance in the genome, however the distance over which linkage disequilibrium is maintained varies widely in different areas of the genome. Thus the length of DNA over which an ideal haplotyping procedure should operate will differ from one gene to another. In general, however, it is desirable to determine haplotypes over distances of at least 2 kb; more preferably at least 5 kb; still more preferably at least 10 kb and most preferably at least 20 kb. Procedures for determining extended haplotypes (i.e. haplotypes >10 kb in length) are emphasized in this application, however, in many cases haplotypes spanning shorter distances may be completely acceptable and may capture all or virtually all of the biologically relevant variation in a larger region of DNA.

In genes that consist of two or more DNA segments that are not in linkage disequilibrium, due to the intervening presence of DNA regions subject to a high frequency of recombination, the preferred approach to haplotype determination is to separately determine haplotypes in each of the two or more constituent regions. The subsequent genetic analysis of genotype - phenotype relationships entails the consideration of all the haplotype groups that exist among the two or more haplotyped segments. Consider, for example, a 15 kb DNA segment in which there is a high frequency of recombination in a central 3 kb segment, but substantial linkage disequilibrium in two flanking 6 kb segments, A and B. The haplotype analysis strategy might consist of determining all the common haplotypes (or haplotype groups - see below) in each of the two 6 kb segments, then considering all the possible combinations of A and B haplotypes. For example if there are three haplotypes or haplotype groups at A (a, a' and a'') and four at B (b, b', b'', b''') then all the combinations (a:b, a:b', a:b'', a:b''', a':b, a':b', a':b'', a':b''', etc.) that occur at, say, a frequency of 5% or greater would be analyzed with respect to relevant phenotypes.

#### *Four Approaches to Haplotyping*

Three groups of haplotyping methods based on allele enrichment are described in this application, plus a fourth group of methods based on visualizing single DNA molecules optically. The usual starting material for all these haplotyping methods is total genomic DNA. In some cases total cellular RNA (or cDNA) may be the starting material. (RNA or cDNA-based methods are predicated on the assumption that both alleles of a gene are transcribed equally; this

assumption does not always hold, therefore it should be tested experimentally in any case where cDNA is being considered as the starting material for a genotyping or haplotyping procedure.)

Each of the three families of allele enrichment methods preferably involves three steps.

The first step is to determine the genotype of at least one polymorphic site in the starting

genomic DNA to provide the basis for design of an allele enrichment procedure. Preferably two or more polymorphic sites are genotyped, and most preferably all polymorphic sites in the DNA segment of interest are genotyped. If two or more sites are heterozygous at the DNA locus of interest then the sample must be subjected to a haplotyping method to determine the haplotypes.

The second step entails obtaining, from a sample of genomic DNA (or RNA or cDNA)

containing two alleles of a gene or other DNA segment of interest, a population of DNA molecules enriched for only one allele. This can be accomplished using any of a variety of novel methods described herein. The third step is a genotyping procedure performed on the enriched DNA. The genotyping procedure will reveal that, at each site which is heterozygous in the subject's genomic DNA (as determined in the first step), only one allele is present in the enriched material. Alternatively, the allele ratio in the enriched material is sufficiently imbalanced, compared to the 1:1 allele ratio in genomic DNA, that the enriched allele can be identified with certainty. In either event, the nucleotides that must be present on the non-enriched allele can be deduced by "subtracting" the haplotype of the enriched allele from the genotype of the starting DNA, determined in step 1. For example, for a DNA segment that is heterozygous at three sites, where site 1 has A or T, site 2 has C or T and site 3 has A or G, if a first haplotype is: 1 = A, 2 = T, 3 = A, then the other haplotype must be: 1 = T, 2 = C, 3 = G. However, rather than simply deducing the second haplotype, a preferred method for haplotype analysis entails the independent determination of both haplotypes present in a sample - by enriching and subsequently genotyping each of the two alleles present in a sample in separate experiments; they should collectively account for the genotype determined from the DNA sample in step one. This practice increases the accuracy of the haplotyping methods described herein.

Step 1 of the procedure described above is mostly dispensable; it is possible to proceed directly to DNA strand enrichment knowing the location of only one polymorphic site (which will provide the basis for designing an enrichment procedure for one allele). Virtually any genotyping procedure will work in step three, however, because allele enrichment is generally not complete, quantitative or semi-quantitative genotyping methods are preferred. Good quantitative genotyping methods will permit accurate haplotypes to be determined even when the degree of allele enrichment is only 2:1, or even less. On the other hand, if substantial allele enrichment is achieved in step two then the genotyping procedure of step three may consist of performing DNA sequencing reactions on the enriched material. For example, chain terminating DNA sequencing reactions could be used to determine the haplotype of the enriched DNA.

A fourth group of haplotyping methods involves microscopic visualisation of single DNA molecules that have been treated in a manner that produces allele specific changes at polymorphic sites. These haplotyping methods are based on the optical mapping and sequencing methods of D. Schwartz, described in US Patent 5,720,928. They are described separately in section 6 below.

### *Three Groups of Allele Enrichment Haplotyping Methods*

The three groups of haplotyping methods dependent on allele enrichment differ in respect to the procedure used to obtain a population of DNA molecules enriched for one of two alleles present in a starting sample. The three different enrichment methods entail: (i) allele capture strategies (summarized in Figure 11 and described in detail below in section 3), applicable to double or single stranded DNA, in which an easily isolated material is linked preferentially to one allele; (ii) allele-selective DNA cleavage strategies, which exploit allele differences in the presence of restriction enzyme cleavage sites. After cleavage of one allele the other allele may be selectively amplified, or separated by a size selection procedure, or the cleaved allele may be removed by an allele selective degradation procedure (summarized in Figure 12 and described below in section 4); and (iii) selective amplification strategies, in which one allele is preferentially amplified from a mixture of two alleles by the design of a primer or primers that exploit sequence differences at polymorphic sites (summarized in Figure 13 and described below in section 5).

### *Degree of allele enrichment required for haplotyping*

Strand enrichment by any of the above methods need not be quantitative or completely selective in order to produce an accurate and reproducible haplotyping result. Even if both alleles are still present after enrichment, as long as one allele is consistently present in greater amount than the other, the enrichment may be adequate to produce a satisfactory discrimination between alleles in a subsequent genotyping test. Preferably the degree of strand enrichment is at least 1.5-fold, more preferably two-fold, more preferably at least four-fold, still more preferably at least six-fold, and most preferably at least ten-fold. Further enrichment beyond 10-fold is desirable, but is unlikely to produce significant changes in the accuracy of the haplotyping test. The adequacy of haplotype determination using a DNA population that is only partially enriched for the desired allele can be determined by repeated analyses of known samples to determine the error rate associated with different known allele ratios.

### *Yield of enriched alleles required for haplotyping*

After allele enrichment, one has a population of DNA molecules for genotyping analysis that is necessarily less than the starting number of DNA molecules because no enrichment procedure will permit 100% recovery of the selected allele. However, just as a high degree of allele selectivity is not necessary during enrichment, a high yield of the enriched allele is not necessary either. The amount of enriched allele will of course depend in part on the quantity of starting DNA. Thus, in a haplotyping experiment that starts with one microgram of genomic DNA, only a small fraction of the alleles in the starting material – as little as 0.1% - have to be captured by the allele enrichment procedure, provided the subsequent genotyping step (usually PCR based) is sensitive enough to amplify an amount of template (~300 copies) that would normally be found in 1 ng of genomic DNA. If necessary the PCR amplification step of the genotyping procedure can be modified to increase sensitivity using methods known in the art, such as nested PCR (two rounds of PCR, first with an outside set of primers, then with an inside set) or an increased number of PCR cycles. Also, to compensate for a low efficiency of captured alleles the quantity of input genomic DNA or cDNA can be increased to 2 ug, 4 ug or even 10 ug or more. Preferably the fraction of input alleles that are captured by the enrichment procedure is at least 0.01% of the starting number of alleles, more preferably at least 0.05%, still more preferably at least .25%, still more preferably at least 2% and most preferably at least 10%. The capture of a still higher fraction of the input alleles does not contribute significantly to the performance of the procedure, and in fact is undesirable if it compromises the selectivity of strand enrichment.

#### *Controlling the size of DNA molecules to be haplotyped*

Before performing allele enrichment procedures on DNA fragments it may be desirable to control the size of the input DNA by random or specific cleavage procedures. One reason is that very long DNA fragments may be significantly more difficult to selectively enrich than shorter fragments (due, for example, to a greater tendency for shear forces to break long fragments, or a greater tendency for long fragments to adhere to or be trapped by particles or matrices required for separation). Therefore it is preferable to produce DNA fragments that are only moderately longer than the size of the region to be haplotyped (which is determined by the biological problem being analyzed, and the location and relationship of DNA polymorphisms, including the degree of linkage disequilibrium in the region being analyzed; see discussion above). The DNA segment to be haplotyped may include a gene, part of a gene, a gene regulatory region such as a promoter, enhancer or silencer element, or any other DNA segment considered likely to play a role in a biological phenomenon of interest.

Production of DNA fragments in the desired size range can be accomplished by using random fragmentation procedures (e.g. shearing DNA physically by pipetting, stirring or by use

of a nebulizer), by partial or complete restriction endonuclease digestion, or by controlled exposure to a DNAase such as *E. coli* DNAase I.

With random or semi-random DNA fragmentation procedures, such as partial nuclease digestion, the aim is to produce a collection of DNA fragments, most of which span the entire region to be haplotyped (and that contain the site that will be used to effect allele enrichment). Mathematical methods can be used to determine the optimal size distribution – for example, a size distribution may be selected in which 80% of the fragments span the target region, assuming random distribution of DNA breakpoints. Preferably at least 50% of the DNA fragments are in this size range.

Complete restriction endonuclease digestion is another useful way to control the size of input DNA molecules, particularly when the full DNA sequence or the restriction map of the DNA segment to be haplotyped is known. Restriction digestion with enzymes that cleave DNA at polymorphic sites produces restriction fragments of different lengths from different alleles (so called restriction fragment length polymorphisms, or RFLPs). Cleaving at restriction sites that produce RFLPs can be used to produce DNA molecules that do or do not contain binding sites for DNA binding molecules (e.g. DNA binding proteins, oligonucleotides, PNAs or small molecules that bind DNA) such that only one of two alleles in a genomic DNA sample contains the binding site. In order for this approach to work the location of all binding sites for the allele specific DNA binding molecule must be taken into account. The preparation of DNA molecules for haplotyping by specific DNA cleavage can be performed so as to produce molecules that will perform optimally in the allele specific binding step.

If single stranded DNA is to be the input material for haplotyping then preferably the optimal size distribution of DNA molecules is obtained while DNA is still double stranded, using any of the methods described above. Subsequently the sample can be denatured, subjected to an allele enrichment step, and subsequently genotyped to determine the haplotypes.

### 3. Haplotyping: Double and single strand-based allele enrichment methods

#### 3.1 Introduction

The goal of allele selection methods is to physically fractionate a genomic DNA sample (the starting material) so as to obtain a population of molecules enriched for one allele of the DNA segment or segments to be analyzed. The details of the procedure depend on the polymorphic nucleotide(s) that provide the basis for allele enrichment and the immediate flanking sequence upstream and/or downstream of the polymorphic site. As explained below, different types of sequence polymorphisms lend themselves to different types of allele enrichment methods.

Once a polymorphic site is selected for allele enrichment the enrichment steps are as follows: (i) prepare DNA fragments for allele enrichment; (ii) add to the DNA fragments a molecule that binds DNA in a sequence specific manner (hereafter referred to as the 'DNA binding molecule') such that one allele of the target DNA segment will be bound and the other not; (iii) allow complexes to form between DNA fragments and the allele specific DNA binding molecule under conditions optimized for allele selective binding; (iv) add a second reagent, such as an antibody, that binds to the allele specific DNA binding molecule (which in turn is bound to DNA fragments, including fragments comprising the selected allele); (v) remove the complex consisting of {selected allele + DNA binding protein + second reagent bound to DNA binding protein} from the starting DNA sample by either physical, affinity (including immunological), chromatographic or other means; (vi) releasing the bound DNA from the complex and (vii) genotyping it to determine the haplotype of the selected allele. These steps are described in greater detail below, except for step (ii), addition of an allele specific DNA binding molecule, which is described at greater length in the following sections, each of which describes one class of allele specific DNA binding molecules in detail.

(i) Preparation of DNA fragments for allele enrichment. The condition of the DNA may be controlled in any of several ways: DNA concentration, size distribution, state of the DNA ends (blunt, 3' overhang, 5' overhang, specific sequence at the end, etc.), degree of elongation, etc. The DNA is also preferably suspended in a buffer that maximizes sequence specific DNA binding. Preferred DNA concentrations for these procedures are in the range from 100 nanograms to 10 micrograms of genomic DNA in a volume of 10 to 1000 microliters. Preferably lower amounts of DNA and lower volumes are used, in order to control costs and to minimize the amount of blood or tissue that must be obtained from a subject to obtain sufficient DNA for a successful haplotyping procedure. As described above, preferably the size of the DNA fragments is controlled to produce a majority of desired fragments which span the DNA segment to be haplotyped. The length of such a segment may vary from 500 nucleotides to 1 kb, 3 kb, 5 kb, 10 kb, 20 kb, 50 kb, 100 kb or more. Fragments of the desired size may be produced by random or specific DNA cleavage procedures, as described above. The exploitation of allele specific sequences at the ends of restriction cleaved DNA molecules for haplotyping is described below. An optimal buffer and binding conditions provide for maximum discrimination between the binding of the allele specific DNA binding molecule to the selected allele vs. the non-selected allele. (The binding of the DNA binding molecule to many other irrelevant DNA fragments in the genomic DNA is unavoidable but should not interfere with the enrichment of the selected allele.)

(ii) Described below are several types of allele specific DNA binding molecules, including proteins, peptides, oligonucleotides, and small molecules as well as combinations of

these molecules. These molecules may be designed or selected to bind double stranded (ds) or single stranded (ss) DNA in a sequence specific manner

(iii) Complexes are formed between DNA and the allele specific DNA binding molecule under conditions optimized for binding specificity - e.g., conditions of ionic strength, pH, temperature and time that promote formation of specific complexes between the binding molecules and the DNA. Optimization of allele selective binding conditions will in general be empirical and, in addition to optimization of salt, pH and temperature may include addition of cofactors. Cofactors include molecules known to affect DNA hybridization properties, such as glycerol, spermidine or tetramethyl ammonium chloride (TMAC), as well as molecules that exclude water such as dextran sulphate and polyethylene glycol (PEG). Optimization of temperature may entail use of a temperature gradient, for example ramping temperature from >95°C down to <40°C.

(iv) After the selected DNA fragment is bound to an allele specific DNA binding molecule a second reagent, such as an antibody, aptamer, streptavidin or nickel coated bead or other ligand that binds to the allele specific DNA binding molecule can be added to the reaction mix. Said second reagent forms complexes with the DNA binding molecules (and any DNA fragments they are bound to) that facilitates their removal from the remaining DNA fragments. This step can be omitted if the DNA binding molecule added in step (ii) already contains or is attached to a ligand or a bead or is otherwise modified in a way that facilitates separation after formation of allele specific complexes in step (iii). For example, if the DNA binding molecule is a protein it can be modified by appending a polyhistidine tag or an epitope for antibody binding such the hemagglutinin (HA) epitope of influenza virus. Then, before, during or after step (iii), nickel coated beads can be added to the DNA sample, or alternatively the sample can be delivered to a nickel column for chromatography, using methods known in the art (e.g. QIAexpress Ni-NTA Protein Purification System, Qiagen, Inc., Valencia, CA). First free DNA is washed through the column, then the DNA bound to the poly-his containing DNA binding protein is eluted with 100 – 200 mM imidazole using methods known in the art. In this way DNA fractions enriched for both alleles (bound and unbound) are collected from one procedure. An equivalent procedure for an epitope tagged DNA binding molecule could include addition of antibody coated beads to form {bead - protein - DNA} complexes which could then be removed by a variety of physical methods (see below). Alternatively the protein - DNA complexes could be run over an antibody column (using an antibody that binds to the epitope engineered into the allele specific DNA binding molecule). An important consideration in designing and optimizing a specific allele enrichment procedure is that the enrichment conditions are sufficiently mild that they do not cause dissociation of the {DNA binding molecule + selected allele} complexes to an

extent that there is too little DNA remaining at the end of the procedure for robust DNA amplification and genotyping.

(v) Separation of complexes containing the {DNA binding molecule + selected allele}, (plus or minus an optional third moiety bound to the DNA binding protein) from the remainder of the DNA sample may be accomplished by physical, affinity (including immunological) or other means. Preferred methods for removing complexes include application of a magnetic field to remove magnetic beads attached to the selected allele via the DNA binding molecule or other moiety; centrifugation (e.g. using a dense bead coated with a ligand like an antibody, nickel, streptavidin or other ligand known in the art, that binds to the DNA binding molecule), or filtration (for example using a filter to arrest beads coated with ligand to which the DNA binding molecule and the attached DNA fragments are bound, while allowing free DNA molecules to pass through), or by affinity methods, such as immunological methods (for example an antibody column that binds the DNA binding molecule which is bound to the selected DNA, or which binds to a ligand which in turn is bound to the DNA binding molecule), or by affinity chromatography (e.g. chromatography over a nickel column if the DNA binding molecule is a protein that has been modified to include a polyhistidine tag, or if the DNA binding molecule is bound to a second molecule that contains such a tag). The separation of the allele specific DNA binding molecule and its bound DNA from the remaining DNA can be accomplished by any of the above or related methods known in the art, many of which are available in kit form from companies such as Qiagen, Novagen, Invitrogen, Stratagene, ProMega, Clontech, Amersham/Pharmacia Biotech, New England Biolabs and others known to those skilled in the art.

(vi) Release the bound DNA from the partially purified complexes containing the selected allele. This step can be accomplished by chemical or thermal denaturing conditions (addition of sodium hydroxide; boiling) or by milder changes in buffer conditions (salt, cofactors) that reduce the affinity of the DNA binding molecule for the selected allele.

(vii) Genotyping the enriched DNA to determine the haplotype of the selected allele can be accomplished by the genotyping methods described herein or by other genotyping methods known in the art, including chemical cleavage methods (Nucleave, Variagenics, Cambridge, MA), primer extension based methods (Orchid, Sequenom, others), cleavase based methods (Third Wave, Madison, WI), bead based methods (Luminex, Illumina) miniaturized electrophoresis methods (Kiva Genetics) or by DNA sequencing. The key requirement of any genotyping method is that it be sufficiently sensitive to detect the amount of DNA remaining after allele enrichment. If there is a small quantity of DNA after allele enrichment (less than 1 nanogram) then it may be necessary increase the number of PCR cycles, or to perform a two step amplification procedure in order to boost the sensitivity of the genotyping procedure. For



example the enriched allele can be subjected to 40 cycles of PCR amplification with a first set of primers, and the product of that PCR can then be subjected to a second round of PCR with two new primers internal to the first set of primers.

No DNA amplification procedure is required in any step of the enrichment procedure until the genotyping step at the end, so allele enrichment methods are not constrained by the limitations of amplification procedures such as PCR. As a result, the length of fragments that can be analyzed is, in principal, quite large. (In contrast, amplification procedures – such as PCR – generally become technically difficult above 5 – 10 kb, and very difficult or impossible above 20 kb, particularly when the template is human genomic DNA or genomic DNA of similar complexity.) It can also be difficult, during amplification (e.g. when using methods such as polymerase chain reaction) to prevent the occurrence of some degree of *in vitro* allele interchange. That is, during denature–renature cycles of the PCR, primer extension products that have not extended all the way to the reverse primer (i.e. incompletely extended strands) may anneal to a different template strand than the one they originated from – in some cases a template corresponding to a different allele – resulting in synthesis of an *in vitro* recombinant DNA product that does not correspond to any naturally occurring allele. In contrast, there is no chance of artifactual DNA strand interchange with the allele enrichment methods described in this application. Apart from these advantages over amplification methods, the strand selection methods described below are also attractive in that the costs of optimizing and carrying out a long range PCR amplification are avoided. Furthermore, the allele enrichment procedures described herein are for the most part generic: the same basic steps can be followed for any DNA fragment.

### 3.2 Double stranded vs. single stranded allele selection methods

Allele selection may be accomplished using single or double stranded DNA. Single stranded DNA is produced by denaturing double stranded DNA – for example by heating or by treatment with alkali, preferably after a sizing procedure has been applied to double stranded DNA to achieve an optimal size distribution of DNA fragments. Both single and double stranded DNA methods have advantages and disadvantages. One advantage of single stranded methods is that the specificity of Watson-Crick base pairing can be exploited for the affinity capture of one allele. Disadvantages of single strand methods include: (i) the propensity of single stranded DNA molecules to anneal to themselves (forming complex secondary structures) or to other, only partially complementary single stranded molecules. For example the ubiquitous human DNA repeat element Alu (which is up to ~280 nucleotides long) may cause two non-complementary strands to anneal; (ii) Single stranded DNA is more susceptible to breakage than double stranded DNA. Strand breaks destroy the physical contiguity that is essential for haplotyping.

Double stranded DNA has several advantages over single stranded DNA as the starting point for the haplotyping methods of this invention. First, it is less susceptible to breakage. Second, it is less likely to bind non-specifically to itself or other DNA molecules (whether single stranded or double stranded). Third, there are a variety of high affinity, sequence specific interactions between double stranded DNA and proteins (e.g. restriction enzymes, transcription factors, natural and artificial zinc finger proteins), as well as high affinity interactions between double stranded DNA and single stranded DNA or modified oligonucleotides (e.g. via Hoogsteen or reverse Hoogsteen base pairing) and between double stranded DNA and small molecules (e.g. polyamides) that can provide the basis for allele enrichment. Another type of structure that can be exploited for allele enrichment is D-loops, formed by strand invasion of a duplex DNA molecule by an oligonucleotide or a DNA-like molecule such as peptide nucleic acid (PNA). D loop formation can be facilitated by addition of *E. Coli* RecA protein, using methods known in the art. Fourth, restriction enzyme cleaved double stranded DNA may have termini that can provide the basis for allele specific treatments, including affinity selection (e.g. ligation to an adapter strand), strand degradation (e.g. allele selective degradation of one allele but not the other), circularization and other procedures described below.

### 3.3 Protein-based allele enrichment methods

In protein-based allele enrichment methods a DNA binding protein with differential affinity for the two allelic DNA segments to be separated is added to a mixture of genomic DNA or cDNA fragments (which have been denatured in the case of single strand methods). Since most sequence specific binding proteins recognize double stranded DNA, dsDNA is the preferred starting material for protein-based allele enrichment methods. Next complexes are formed between the binding protein and DNA segments containing the sequence motif recognized by the binding protein. The DNA segments to be haplotyped must differ in respect to the presence or absence of the protein binding site. As described above, this requirement can be met in several ways, most preferably by using restriction endonuclease(s) to produce DNA fragments that contain only one binding site (selected allele), or none (other allele), for the binding protein. Subsequently the protein – DNA complexes are purified from the DNA that is not protein bound. This purification can be accomplished by adding a second reagent that (i) binds to the binding protein, and (ii) can be physically separated from the mixture. The second reagent may be a bead, a magnetic particle, a surface or any other structure that facilitates the physical separation step. It may be modified by coupling to a protein binding reagent. Specific reagents for binding to proteins include antibodies, nickel-polyhistidine affinity, avidin-biotin affinity and similar schemes known to those skilled in the art. The physical separation of the complexes containing binding protein bound to DNA can be effected by gravity, centrifugation, a magnetic field,

filtration, chromatography, electrophoresis or other means. The final step is to genotype the purified material, which yields the haplotype.

### 3.3.1 Protein-based double stranded allele selection methods

5 The major categories of naturally occurring sequence specific DNA binding proteins include zinc finger proteins and helix-turn-helix transcription factors. In addition, proteins that normally act on DNA as a substrate can be made to act as DNA binding proteins either by (i) alterations of the aqueous environment (e.g. removal of ions, substrates or cofactors essential for the enzymatic function of the protein, such as divalent cations) or (ii) by mutagenesis of the protein to disrupt catalytic, but not binding, function. Classes of enzymes that bind to specific dsDNA sequences include restriction endonucleases and DNA methylases. (For a recent review see: Roberts R.J. and D. Macelis. REBASE - restriction enzymes and methylases. *Nucleic Acids Res.* 2000 Jan 1;28(1):306-7.) Finally, *in vitro* evolution methods (DNA shuffling, dirty PCR and related methods) can be used to create and select proteins or peptides with novel DNA binding properties. The starting material for such methods can be the DNA sequence of a known DNA binding protein or proteins, which can be mutagenized globally or in specific segments known to affect DNA binding, or can be otherwise permuted and then tested or selected for DNA binding properties. Alternatively the starting material may be genes that encode enzymes for which DNA is a substrate - e.g. restriction enzymes, DNA or RNA polymerases, DNA or RNA helicases, topoisomerases, gyrases or other enzymes. Such experiments might be useful for producing sequence specific ssDNA binding proteins, as well as sequence specific dsDNA binding proteins. For recent descriptions of *in vitro* evolution methods see: Minshull J. and W.P. Stemmer: Protein evolution by molecular breeding. *Curr Opin Chem Biol.* 1999 Jun;3(3):284-90; Giver, L., and F.H. Arnold: Combinatorial protein design by *in vitro* recombination. *Curr Opin Chem Biol.* 1998 Jun;2(3):335-8; Bogarad, L.D. and M.W. Deem: A hierarchical approach to protein molecular evolution. *Proc Natl Acad Sci U S A.* 1999 Mar 16;96(6):2591-5; Gorse, D., Rees, A., Kaczorek, M. and R. Lahana: Molecular diversity and its analysis. *Drug Discov Today.* 1999 Jun;4(6):257-264.

Among the classes of DNA binding proteins enumerated above which could be used to select DNA molecules, a preferred class of proteins would have the following properties: (i) any two sequences differing by one nucleotide (or by one nucleotide pair in the case of dsDNA) could be discriminated, not limited by whether or not one version of the sequence is a palindrome, or by any other sequence constraint, (ii) DNA binding proteins can be designed or selected using standard conditions, so that the design or selection of proteins for many different sequence pairs is not onerous. (This requirement arises from the concern that, in order to be able to readily select any given DNA molecule for haplotyping it is desirable to have a large collection

of DNA binding proteins, each capable of discriminating a different pair of sequences.) (iii) The affinity of the protein for the selected DNA sequence is sufficient to withstand the physical and/or chemical stresses introduced in the allele enrichment procedure. (iv) The DNA binding molecules are stable enough to remain in native conformation during the allele enrichment procedure, and can be stored for long periods of time. (v) The length of sequence bound by the allele specific DNA binding protein is preferably at least six nucleotides (or nucleotide pairs), more preferably at least 8 nucleotides, and most preferably 9 nucleotides or longer. The longer the recognition sequence, the fewer molecules in the genomic DNA fragment mixture will be bound, and therefore the less 'background' DNA there will be accompanying the enriched allele. In addition to the five foregoing criteria, it may be desirable to make a fusion between the DNA binding protein and a second protein so as to facilitate enrichment of the DNA binding protein. For example, appending an epitope containing protein would allow selection by antibody based methods. Appending six or more histidine residues would allow selection by zinc affinity methods. (DNA binding proteins may also be useful in microscopy-based haplotyping methods described elsewhere in the application, and for that purpose it may be useful to make a fusion with a protein that produces a detectable signal - for example green fibrillary protein.)

### 3.3.2 Zinc finger proteins

Given the above criteria, zinc finger proteins are a preferred class of DNA binding proteins. It is well established that zinc finger proteins can bind to virtually any DNA sequence motif; in particular, they are not limited to pallindromic sequences, as both type II restriction endonucleases and helix-turn-helix transcription factors are. See, for example: Choo, Y. and A. Klug (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91: 11163-11167. Jamieson, A.C., Wang, H. and S.-H. Kim. (1996) A Zinc finger directory for high-affinity DNA recognition. *Proc. Natl. Acad. Sci. U. S. A.* 93: 12834 -12839. Segal, D.J., Dreier, B., Beerli, R.R. and C.F. Barbas (1999) Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. *Proc. Natl. Acad. Sci. U. S. A.* 96: 2758-2763. Segal, D.J. and C.F. Barbas (2000) Design of novel sequence specific DNA-binding proteins. *Curr. Opin. Chem. Biol.* 4: 34-39. These papers and other work in the field demonstrate that it is possible to generate zinc finger proteins that will bind virtually any DNA sequence from 3 nucleotides up to 18 nucleotides. Further, these studies show that *in vitro* generated zinc finger proteins are capable of binding specific DNA sequences with low nanomolar or even subnanomolar affinity, and are capable of distinguishing sequences that differ by only one base pair with 10 to 100-fold or even greater differences in affinity. It has also been demonstrated that zinc finger proteins can be modified by fusion with other protein domains that provide detectable labels or attachment domains. For example zinc finger proteins can be fused with

jellyfish green fibrillary protein (GFP) for labelling purposes, or fused to polyhistidine at the amino or carboxyl terminus, or fused with an antibody binding domain such as glutathione transferase (GST) or influenza virus hemagglutinin (HA) (for which there are commercially available antisera) for attachment and selection purposes.

5           Methods for making zinc finger proteins of desired sequence specificity are well known in the art and have recently been adapted to large scale experiments. See, in addition to the above references: Beerli R.R., Dreier B. and C.F. Barbas (2000) Positive and negative regulation of endogenous genes by designed transcription factors. *Proc Natl Acad Sci U S A.* 97: 1495-1500; Beerli R.R., Segal D.J., Dreier B. and C.F. Barbas (1998) Toward controlling gene  
10       expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. *Proc Natl Acad Sci U S A.* 95: 14628-14633.) Methods for using phage display to select zinc finger proteins with desired specificity from large libraries have also been described: Rebar E.J. and C.O. Pabo (1994) Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science*.  
263(5147):671-673. Rebar E.J., Greisman H.A. and C.O. Pabo (1996) Phage display methods for selecting zinc finger proteins with novel DNA-binding specificities. *Methods Enzymol.* 267:129-149.) The phage display method offers one way to bind selected alleles to a large complex that can be efficiently removed from a mixture of DNA fragments. Preventing nonspecific DNA binding to intact phage requires careful optimization of blocking conditions.

20           For the haplotyping methods described in this application the length of the DNA sequence recognized by a zinc finger protein may range from 3 nucleotides to 18 or more nucleotides. Sequences less than 6 nucleotides are generally not useful for haplotyping because, unless they contain the dinucleotide CpG, they occur too frequently in DNA to allow haplotyping over distances of several kb or longer - that is, for DNA fragments greater than, say, 3 kb, both  
25       alleles will frequently have one or more copies of any sequence motif that is shorter than six nucleotides, because the average interval between three, four and five nucleotide repeats (assuming random order sequence and equal occurrence of each of the four nucleotides, neither of which actually obtains) is  $4 \times 3 = 64$ ,  $4 \times 4 = 256$  and  $4 \times 5 = 1024$ . If both alleles have a copy of a target sequence then clearly that sequence cannot be used to selectively enrich one  
30       allele. Preferred zinc finger proteins recognize 6, 9, 12 or 18 nucleotides, with the longer sequences preferred. However, the length of sequence bound is only one of several important considerations in selecting an optimal zinc finger protein. Equally important are the specificity of binding and the affinity of binding. Preferably a zinc finger protein has a specificity of at least 10 fold, and more preferably 100 fold or greater, with respect to all sequences that differ from the  
35       selected sequence by one or more nucleotides. This level of specificity will enable the allele selectivity required for successful allele enrichment. Optimal zinc finger proteins must also have

a high affinity for the selected sequence. Preferably the dissociation constant of the zinc finger protein for the target DNA sequence is less than 50 nanomolar, more preferably less than 10 nanomolar, and most preferably less than 2 nanomolar. Multiple zinc finger proteins that meet all the enumerated criteria have been produced, demonstrating the ability of the skilled artisan to accomplish the necessary modifications to naturally occurring zinc finger proteins. Methods for improving the specificity and affinity of binding include random or site directed mutagenesis, selection of phage bearing mutant zinc finger proteins with desired specificity from large libraries of phage, and *in vitro* evolution methods.

Because each zinc finger recognizes three nucleotides, one way to make zinc finger proteins that recognize sequences of six nucleotides or longer is to assemble two or more zinc fingers with known binding properties. The use of zinc fingers as modular building blocks has been demonstrated by Barbas and colleagues (see: *Proc Natl Acad Sci U S A.* 95: 14628-14633, 1998) for nucleotide sequences of the form (GNN)<sub>x</sub> where G is guanine, N is any of the four nucleotides, and x indicates the number of times the GNN motif is repeated.

A large number of zinc finger proteins exist in nature, and a still larger number have been created *in vitro* (e.g. see work cited above). Any of these known zinc finger proteins may constitute a useful starting point for the construction of a useful set of allele specific DNA binding proteins. The protein Zif268 is the most extensively characterized zinc finger protein, and has the additional advantage that there is relatively little target site overlap between adjacent zinc fingers, making it well suited to the modular construction of zinc finger proteins with desired DNA sequence binding specificity. See, for example: Segal, D.J., et al. *Proc Natl Acad Sci U S A.* 96: 2758-2763, 1999. Zif268 is a preferred backbone for production of mutant zinc finger proteins.

### 3.3.3 Restriction endonucleases

Another class of sequence specific DNA binding proteins useful for allele enrichment is restriction endonucleases. There are over 400 commercially available restriction endonucleases, and hundreds more that have been discovered and characterized with respect to their binding specificity. (Roberts R.J. and D. Macelis. *Nucleic Acids Res.* 2000 Jan 1;28(1):306-7.)

Collectively these enzymes recognize a substantial fraction of all 4, 5 and 6 nucleotide sequences (of which there are 256, 1024 and 4096, respectively). For certain polymorphic nucleotides, the exquisite sequence specificity of these enzymes can be used to selectively bind one allelic DNA fragment that contains the cognate recognition site, while not binding to the DNA fragment corresponding to the other allele, which lacks the cognate site. Restriction endonucleases do not have the flexibility that zinc finger proteins do in being able to selectively bind virtually any sequence up to 18 nucleotides in length, but they do have the attraction of being highly specific,

readily available, and for the most part inexpensive to produce. The identification of polymorphic sites that lie within restriction enzyme binding sequences will become much simpler as the sequence of the human genome is completed, and the generation of restriction maps becomes primarily a computational, rather than an experimental, activity.

5 In order for restriction endonucleases to be useful as DNA binding proteins their DNA cleaving function must first be neutralized or inactivated; otherwise DNA is cleaved and released. Inactivation can be accomplished in two ways. First, one can add restriction endonucleases to DNA, allow them to bind under conditions that do not permit cleavage, and then remove the DNA-protein complex. The simplest way to prevent restriction enzyme  
10 cleavage is to withhold divalent cations from the buffer. Second, one can alter restriction endonucleases so that they still bind DNA but can not cleave it. This can be accomplished by altering the sequence of the gene encoding the restriction endonuclease, using methods known in the art, or it can be accomplished by post-translational modification of the restriction endonuclease, using chemically reactive small molecules.

The first approach – withholding essential cofactors, such as magnesium or manganese - has the advantage that no modification of restriction enzymes or the genes that encode them is necessary. Instead, conditions are determined that permissive for binding but nonpermissive for cleavage.

It may not be possible to identify such conditions for all restriction enzymes; some  
20 enzymes appear to require divalent cations for specific, high affinity recognition of cognate DNA. For such enzymes it may be possible to produce mutant forms that do not require divalent cations for high affinity, specific binding to cognate DNA. For example, mutants of the restriction enzyme Mun I (which binds the sequence CAATTG) have been produced that recognize and bind (but do not restrict) cognate DNA with high specificity and affinity in the  
25 absence of magnesium ion. In contrast, wild type Mun I does not exhibit sequence specific DNA binding in the absence of magnesium ion. The amino acid changes in the mutant Mun I enzymes (D83A, E98A) have been proposed to simulate the effect of magnesium ion in conferring specificity. See, for example: Lagunavicius, A. and V. Siksnys (1997) Site-directed mutagenesis of putative active site residues of Mun I restriction endonuclease: replacement of catalytically  
30 essential carboxylate residues triggers DNA binding specificity. *Biochemistry* 36: 11086-11092.

Structural modification of restriction enzymes to alter their cleaving properties but not their binding properties in the presence of magnesium ion has been also been demonstrated. For example, in studies of the restriction enzyme Eco R I (which binds the sequence GAATTC) it has been demonstrated that DNA sequence recognition and cleaving activity can be dissociated.

35 Studies have shown that mutant Eco RI enzymes with various amino acid substitutions at residues Met137 and Ile197 bind cognate DNA (i.e. 5' – GAATTC – 3') with high specificity but

cleave with reduced or unmeasurably low activity. See: Ivanenko, T., Heitman, J. and A. Kiss (1998) Mutational analysis of the function of Met137 and Ile197, two amino acids implicated in sequence specific DNA recognition by the Eco RI endonuclease. *Biol. Chem.* 379: 459-465.

Other work has led to the identification of mutant Eco RI proteins that have substantially

5 increased affinity for the cognate binding site, while lacking cleavage activity. For example, the Eco RI mutant Gln111 binds GAATTC with ~1,000 fold higher affinity than wild type enzyme, but has ~10,000 lower rate constant for cleavage. (See: King, K., Benkovic, S.J. and P. Modrich [1989] Glu-111 is required for activation of the DNA cleavage center of EcoRI endonuclease *J. Biol. Chem.* 264: 11807-15.) Eco RI Gln111 has been used to image Eco RI sites in linearized  
10 3.2 – 6.8 kb plasmids using atomic force microscopy, a method that exploits the high binding affinity and negligible cleavage activity of the mutant protein. The Eco RI Gln111 protein is a preferred reagent for the methods of this invention, as a reagent for the selective enrichment of alleles that contain a GAATTC sequence (and consequent depletion of alleles that lack such a sequence). Exemplary conditions for selective binding of Eco RI Gln111 to DNA fragments with cognate sequence may include ~50 - 100 mM sodium chloride, 10 – 20 mM magnesium ion (e.g. MgCl<sub>2</sub>) and pH 7.5 in tris or phosphate buffer. Preferably there is molar equivalence of Eco RI Gln111 and cognate DNA binding sites in the sample (e.g. genomic DNA); more preferably there is a 5, 10, 20 or 50 - fold molar excess of enzyme over DNA. Preferred methods for enrichment of the Eco RI bound allele from the non - bound allele include the synthesis of a fusion protein between Eco RI Gln111 and a protein domain that includes an antibody binding site for a commercially available enzyme. Influenza hemagglutinin, beta galactosidase or glutathione S transferase and polyhistidine domains are available as commercial kits for protein purification.

There are several schemes for producing, from genomic DNA, two homologous (allelic)  
25 fragments of a gene that differ in respect to the presence or absence of a sequence such as an Eco RI site. Scheme 1: if the complete sequence of the region being haplotyped is known then the location and identity of all restriction sites, including the subset of restriction sites that include polymorphic nucleotides in their recognition sequence, can be determined trivially by computational analysis using commercially available software. Those restriction sites that  
30 overlap polymorphic nucleotides in the DNA segment of interest can be assessed for suitability as allele enrichment sites. The optimal characteristics of an allele enrichment site include: (i) The site occurs once, or not at all (depending on the allele) in a DNA segment to be haplotyped. This is crucial since the basis of the allele enrichment is the attachment of a protein to the binding site in the allele to be enriched, and its absence in the other allele present in the genomic  
35 DNA sample being haplotyped. (ii) There is a pair of nonpolymorphic restriction sites, different



from the site being used for allele enrichment, that flank the polymorphic site and span a DNA segment deemed useful for haplotype analysis.

The steps for allele enrichment then comprise: restrict genomic DNA with the selected enzyme(s) that flank the polymorphic site so as to produce a DNA segment useful for haplotype analysis (as well as many other genomic DNA fragments); add the DNA binding protein (i.e. the cleavage-inactive restriction enzyme) in a buffer that promotes specific binding to the cognate site (and, if necessary, prevents the restriction enzyme from cleaving its cognate site); selectively remove the restriction enzyme – complex from the genomic DNA by any of the physical or affinity based methods described above – antibody, nickel – histidine, etc. Subsequently, suspend the enriched allele in aqueous buffer and genotype two or more polymorphic sites to determine a haplotype. Scheme 2 is similar but does not require a specific restriction step. Instead, one randomly fragments genomic DNA into segments that, on average, are approximately the length of the segment to be haplotyped. Then add the DNA binding protein and proceed with the enrichment as above. The disadvantage of this scheme is that there may be DNA fragments that include non-polymorphic copies of the cognate sequence for the DNA binding protein. The presence of such fragments will limit the degree of allele enrichment because they will co-purify with the targeted allele, and produce background signal in the subsequent analysis steps. This problem can be addressed by reducing the average size of the fragments in the random fragmentation procedure.

Because of the requirement that the enriched allele fragment have zero or one copies of the sequence to be used for attachment of the restriction, optimal restriction enzymes for these haplotyping methods recognize sequences of 5 nucleotides or greater; preferably they recognize sequence of 6 nucleotides or greater; preferably the cognate sites of such enzymes contain one or more dinucleotides or other sequence motifs that are proportionately underrepresented in genomic DNA of the organism that is being haplotyped; preferably, for haplotyping methods applied to mammalian genomic DNA, they contain one or more 5'-CpG-3' sequences, because CpG dinucleotides are substantially depleted in mammalian genomes. Restriction enzymes that include CpG dinucleotides include Taq I, Msp I, Hha I and others known in the art.

#### 3.3.4 Additional restriction endonuclease based methods

A limitation of the restriction enzyme based allele capture method is that the length of DNA fragment that can be haplotyped depends on the local restriction map. In some cases it may be difficult to find a polymorphic restriction site for which a cleavage-inactive restriction enzyme is available *and* for which the nearest 5' and 3' flanking sequences are at an optimal distance for haplotyping; often the flanking restriction enzyme cleavage sites will be closer to the polymorphic site than desired, limiting the length of DNA segment that can be haplotyped. For

example, it may be optimal from a genetic point of view to haplotype a 15 kb segment of DNA, but there may be no polymorphic restriction sites that are flanked by sites that allow isolation of the desired 15 kb segment. One approach to this problem is to haplotype several small DNA fragments that collectively span the 15 kb segment of interest. A composite haplotype can then be assembled by analysis of the overlaps between the small fragments.

A more general, and more useful, method for circumventing the limitations occasionally imposed by difficult restriction maps is to incorporate aspects of the RecA assisted restriction endonuclease (RARE) method in the haplotyping procedure. (For a description of the RARE procedure see: Ferrin, L.J. and R.D. Camerini-Otero [1991] *Science* 254: 1494-1497; Koob, M. et al. [1992] *Nucleic Acids Research* 20: 5831-5836.) When the RARE techniques are used in the protein mediated allele enrichment method it is possible to haplotype DNA segments of virtually any length, regardless of the local restriction site map.

First, the DNA is sized, either by random fragmentation to produce fragments in the right size range (e.g. approximately 15 kb average size), or one can use any restriction endonuclease or pair of restriction endonucleases to cleave genomic DNA (based on the known restriction map) so as to produce fragments spanning the segment to be haplotyped. In the RARE haplotyping procedure one then uses an oligonucleotide to form a D loop with the segment of DNA that contains the polymorphic restriction site (the site that will ultimately be used to capture the DNA segment to be haplotyped). (The other copy of the allele present in the analyte sample lacks the restriction enzyme sequence as a consequence of the polymorphism.) Formation of the D loop can be enhanced by addition of *E. Coli* RecA protein, which assembles around the single stranded DNA to form a nucleoprotein filament which then slides along double stranded DNA fragments until it reaches a complementary strand. RecA protein, in a complex with a gamma-S analog of ATP and a 30-60 nucleotide long oligodeoxynucleotide complementary or identical to the sequence-targeted site in which the protected restriction site is embedded, then mediates strand invasion by the oligodeoxynucleotide, forming the D loop.

Once this loop is formed the next step is to methylate all copies of the polymorphic restriction site using a DNA methylase. Substantially all copies of the restriction site present in the genomic DNA mixture are methylated. (One nucleotide, usually C, is methylated.) The one polymorphic restriction site which participates in the D loop is not methylated because the D loop is not recognized by the DNA methylase. Next the D loop is disassembled and the methylase inactivated or removed. This leaves the targeted restriction site available for restriction enzyme binding (on the one allele that contains the restriction site). Finally, the restriction-inactive but high affinity binding protein (e.g. Eco RI Gln111) is added to the mixture of genomic DNA fragments. The only fragment that should have an available Eco RI site is the fragment to be haplotyped. Any of several methods can be used to selectively remove that

fragment: the cleavage-inactive restriction enzyme can be fused to a protein that serves as a handle to facilitate easy removal by nickel-histidine, antibody-antigen or other protein-protein interaction, as described in detail elsewhere in this invention. Alternatively, an antibody against the restriction enzyme can be prepared and used to capture the restriction enzyme - allele  
 5 fragment complex to a bead or column to which the antibody is bound, or other methods known in the art can be employed.

The advantage of the RARE assisted haplotyping method is that the local restriction map, and in particular the occurrence of other Eco RI sites (in this example) nearby, is no longer a limitation. Further, the methylation of all sites save the polymorphic site eliminates the  
 10 preference for restriction enzymes that recognize 6 or more nucleotides. With the RARE haplotyping technique any enzyme, including one that recognizes a four nucleotide sequence, is effective for allele enrichment. This is a particularly useful aspect of the invention because four nucleotide sequences recognized by restriction enzymes more often encompass polymorphic sites than 5 or 6 nucleotide sequences, and there are more DNA methylases for 4 nucleotide sequences than for 6 nucleotide sequences recognized by restriction enzymes. Preferred restriction sites for RARE assisted haplotyping are those for which DNA methylases are commercially available, including, without limitation, Alu I, Bam HI, Hae III, Hpa II, Taq I, Msp I, Hha I, Mbo I and Eco RI methylases.

The use of peptides for allele enrichment is described below in the discussion of small molecules that can be used for allele enrichment.

### 3. 4 Haplotype determination by nucleic acid-based allele selection methods

In another aspect of the invention, nucleic acids and nucleic acid analogs that bind specifically to double stranded DNA can be targeted to polymorphic sites and used as the basis  
 25 for physical separation of alleles. Ligands attached to the targeting oligonucleotides, such as, without limitation, biotin, fluorescein, polyhistidine or magnetic beads, can provide the basis for subsequent enrichment of bound alleles. Sequence specific methods for the capture of double stranded DNA, useful for the haplotyping methods of this invention, include: (i) Triple helical interactions between single stranded DNA (e.g. oligonucleotides) and double stranded DNA via  
 30 Hoogsteen or reverse Hoogsteen base pairing. (ii) D-loop formation, again between a single stranded DNA and a double stranded DNA. (iii) D-loop formation between peptide nucleic acid (PNA) and a double stranded DNA. (iv) *in vitro* nucleic acid evolution methods (referred to as SELEX) can be used to derive natural or modified nucleic acids (aptamers) that bind double stranded DNA in a sequence specific manner via any combination of Watson-Crick or Hoogsteen  
 35 base pairing, hydrogen bonds, van der Waals forces or other interaction.

The D loop is formed by the displacement of one strand of the double helix by the invading single strand. RecA protein, as indicated above, facilitates D Loop formation, albeit with only limited stringency for the extent of homology between the invading and invaded sequences. nucleotide analogs Such interactions are useful for allele enrichment when a polymorphic site lies in a sequence context that conforms to the requirements for Hoogsteen or reverse Hoogsteen base pairing. The sequence requirements generally include a homopyrimidine/homopurine sequence in the double stranded DNA, however the discovery of nucleotide analogs that base pair with pyrimidines in triplex structures has increased the repertoire of sequences which can participate in triple stranded complexes. Nonetheless, more general scheme for selective binding to polymorphic DNA sequences is preferable.

### 3 (b). Haplotype determination by nucleic acid-based double stranded allele selection methods

In another aspect of the invention, nucleic acids that bind specifically to double stranded DNA can be targeted to polymorphic sites and used as the basis for physical separation of alleles. The best known types of specific interactions involve triple helical interactions formed via Hoogsteen or reverse Hoogsteen base pairing. These interactions are useful for haplotyping when a polymorphic site lies within a sequence context that conforms to the requirements for Hoogsteen or reverse Hoogsteen base pairing. These requirements typically include a homopyrimidine/homopurine sequence, however the discovery of nucleic acid modifications that permit novel base pairings is resulting in an expanded repertoire of sequences. Nonetheless, a more general scheme for selective binding to polymorphic DNA sequences is preferable.

In another aspect of the invention the formation of D loops by strand invasion of dsDNA can be the basis for an allele specific interaction, and secondarily for an allele enrichment scheme. Peptide nucleic acid (PNA) is a preferred material for strand invasion. Due to its high affinity DNA binding PNA has been shown capable of high efficiency strand invasion of duplex DNA. (Peffer NJ, Hanvey JC, Bisi JE, et al. Strand-invasion of duplex DNA by peptide nucleic acid oligomers. *Proc Natl Acad Sci U S A*. 1993 Nov 15;90(22): 10648-52; Kurakin A, Larsen HJ, Nielsen PE. Cooperative strand displacement by peptide nucleic acid (PNA). *Chem Biol*. 1998 Feb;5(2):81-9. The basis of a PNA strand invasion affinity selection would be conceptually similar to protein-based methods, except the sequence-specific DNA-PNA complexes formed by strand invasion are the basis of an enrichment procedure that exploits an affinity tag attached to the PNA. The affinity tags may be a binding site for an antibody such as fluorescein or rhodamine, or polyhistidine (to be selected by nickel affinity chromatography), or biotin, (to be selected using avidin- or streptavidin-coated beads or surface) or other affinity selection schemes known to those skilled in the art.

In another embodiment of the invention, *in vitro* nucleic acid evolution methods (referred to as aptamers or SELEX) can be used to derive natural or modified nucleic acids that bind double stranded DNA in a sequence specific manner. Methods for high throughput derivation of nucleic acids capable of binding virtually any target molecule have been described. (Drolet DW, 5 Jenison RD, Smith et al. A high throughput platform for systematic evolution of ligands by exponential enrichment (SELEX). *Comb Chem High Throughput Screen*. 1999 Oct;2(5):271-8.)

### 3 (c). Other double stranded allele selection methods

In another aspect of the invention, non-protein, non-nucleic acid molecules can be the 10 basis for affinity selection of double stranded DNA. (Mapp AK, Ansari AZ, Ptashne M, et al. Activation of gene expression by small molecule transcription factors. *Proc Natl Acad Sci U S A*. 2000 Apr 11;97(8):3930-5; Dervan PB, Burli RW. Sequence-specific DNA recognition by polyamides. *Curr Opin Chem Biol*. 1999 Dec;3(6):688-93; White S, Szewczyk JW, Turner JM, et al. Recognition of the four Watson-Crick base pairs in the DNA minor groove by synthetic ligands. *Nature*. 1998 Jan 29;391(6666):468-71.)

## 4. Haplotype determination by allele specific capture of single stranded DNA or RNA

In this example, modified oligonucleotides or modified nucleotide triphosphates are used as affinity reagents to partially purify a complementary DNA species (the allele to be haplotyped) with which they have formed a duplex. The nucleotide or oligonucleotide modification may constitute, for example, addition of a compound that binds with high affinity to a known partner – such as biotin/avidin or polyhistidine/nickel – or it may consist of covalent addition of a compound for which high affinity antibodies are available – such as rhodamine or fluorescein – or it may consist of addition of a metal that allows physical separation using a magnetic field, or 25 it may involve addition of a reactive chemical group that, upon addition of a specific reagent or physical energy (e.g. uv light) will form a covalent bond with a second compound that in turn is linked to a molecule or structure that enables physical separation.

An example of such a modification would be biotin. DNA or RNA once hybridized to biotinylated oligonucleotides or nucleotides could be separated from non-hybridized DNA or 30 RNA using streptavidin on a solid support. Other possible modifications could include but are not limited to: antigens and antibodies, peptides, nucleic acids, and proteins that when attached to oligonucleotides or nucleotides would bind to some other molecule on a solid support. For the purposes of this disclosure, biotin will be used as the exemplary modification of the oligonucleotides or nucleotides and streptavidin will be attached to the solid support.

35 Oligonucleotides used in the descriptions below can be comprised of either normal nucleotides and/or linkages or modified nucleotides and/or linkages. The only requirement is

that the oligonucleotides retain the ability to hybridize DNA or RNA and that they can be utilized by the appropriate enzymes if necessary. Examples of modified oligonucleotides could include but are not limited to: peptide nucleic acid molecules, phosphorothioate and methylphosphonate modifications. The term oligonucleotide when used below will refer to both natural and modified oligonucleotides.

The following are examples for employing allele specific capture of DNA or RNA to determine haplotypes:

1. A biotinylated oligonucleotide directed against a site that is heterozygous for a nucleotide variance, is allowed to hybridize to the target DNA or RNA under conditions that will result in binding of the oligonucleotide to only one of the two alleles present in the sample. The length, the position of mismatch between the oligonucleotide and the target sequence, and the chemical make-up of the oligonucleotide could all be adjusted to maximize the allele specific discrimination. Streptavidin on a solid support is used to remove the biotinylated oligonucleotide and any DNA or RNA associated by hybridization to the oligonucleotide. For example, allele 1 is specifically captured by hybridization of an oligonucleotide containing a T at the variance site. The target DNA or RNA from allele 1 is then disassociated from the primer and solid support under denaturing conditions. The isolated RNA or DNA from allele 1 can then be genotyped to determine the haplotype. Alternatively, the RNA or DNA remaining in the sample, allele 2, following capture and removal of allele 1 could be genotyped to determine the its haplotype.

2. The target DNA is incubated with two oligonucleotides, one of which is biotinylated (Figure 15). If RNA is to be used in this example it must first be converted to cDNA. The oligonucleotides are designed to hybridize adjacent to one another at the site of variance. For example, the 3' end of the biotinylated oligonucleotide hybridizes one base 5' of the variant base. The other oligonucleotide hybridizes adjacent to the biotinylated primer with the 5' most oligonucleotide hybridizing to the variant base. If there is a perfect match at the site of variance (allele 1), the two primers are ligated together. However, if there is a mismatch at the site of variance (allele 2) no ligation occurs. The sample is then allowed to bind to the streptavidin on the solid support under conditions which are permissive for the hybridization of the ligated oligonucleotides but non-permissive for the hybridization of the shorter non-ligated oligonucleotides. The captured oligonucleotides and hybridized target DNA are removed from the sample, the target DNA eluted from the solid support, and genotyped to determine haplotype.

Alternatively, the allele 2 can be genotyped to determine haplotype after removal of allele 1 from the sample.

The size of the oligonucleotides can be varied in order to increase the likelihood that hybridization and ligation will only occur when the correct allele is present. The ligation can be done under conditions which will only allow the hybridization of a shorter oligonucleotide if it is hybridized next to the perfectly matched oligonucleotide and can make use of the stacking energy for stabilization. Also, either the biotinylated oligonucleotide or the other oligonucleotide can contain the mismatch. The biotin can also be put on the 5' or 3' end of the oligonucleotide as long as it is not at the site of ligation.

3. An oligonucleotide is hybridized to the target DNA in which the 3' end of the oligonucleotide is just 5' of the variant base. If RNA is to be used in this example it is first converted to cDNA. The sample is then incubated in the presence of four dideoxy nucleotides with a polymerase capable of extending the primer by incorporating dideoxy nucleotides where one of the dideoxy nucleotides contains a biotin. The biotinylated dideoxy nucleotide is selected to correspond to one of the variant bases such that it will be incorporated only if the correct base is at the site of variance. For example, the base chosen is biotin ddTTP which will be incorporated only when the primer anneals to allele 1. The primer with the incorporated biotinylated dideoxy nucleotide hybridized to allele 1 is separated from the rest of the DNA in the sample using streptavidin on a solid support. The isolated allele 1 can then be eluted from the solid support and genotyped to determine haplotype. As above, allele 2 which is left in the sample after capture and removal of allele 1, can also be genotyped to determine haplotype.

The dideoxy and biotinylated nucleotide do not have to be the same nucleotide. The primer could be extended in the presence of one biotinylated nucleotide, one dideoxy nucleotide and two normal nucleotides. For example, a biotinylated dTTP and a normal dGTP would be added in with another normal nucleotide (not dTTP or dGTP) and a dideoxy nucleotide (not ddTTP or ddGTP). The dideoxy nucleotide would be chosen so that the extension reaction would be terminated before the occurrence of another site for the incorporation of the biotinylated dTTP. Extension from the primer on allele 1 would result in the incorporation of a biotinylated dTTP. Extension from the primer on allele 2 would result in the incorporation of a normal dGTP. Streptavidin on a solid support could be used to separate allele 1 from allele 2 for genotyping to determine haplotype.

#### 4 (a). Protein-based single stranded allele selection methods

## 4 (c). Other single stranded allele selection methods

## 5. Post-allele selection genotyping methods

5

**6. Allele Selective Amplification Haplotyping Methods**

Described in this application is an alternative method for obtaining allele specific PCR products. Our method also entails using modified primers, however the basis for achieving allele specific amplification is the formation of a duplex or secondary structure involving base pairing between (i) nucleotides at or near the 3' end of a strand (said nucleotides being at least partially templated by a primer for the complementary strand) and (ii) nucleotides of the same strand that lie further interior from the 3' end and include (crucially) a polymorphic site (or sites), such that: (i) the secondary structure is formed to a different extent in the two alleles (ideally the secondary structure is formed in a completely allele specific manner), and (ii) the secondary structure at least partially inhibits primer binding and/or primer extension, and consequently inhibits amplification of the strand with the secondary structure at the 3' end. The point of the primer modification, then, is to produce a template for polymerization on the complementary strand leading to a sequence that will form a secondary structure that will inhibit further primer binding/extension from that end. The modification in the primer can be introduced either at the 5' end or internally, but not at the 3' end of the primer. An example of this method applied to haplotyping the ApoE gene is provided below (Example 3), along with Figures 14-17 that illustrate some of the types of secondary structure that can be produced to inhibit primer binding/extension.

One implementation of the method entails introducing a 5' extension in a primer. After a complementary strand is extended across that primer, and then separated by a cycle of denaturation, the complementary strand forms a hairpin loop structure in one allele but not the other. Specifically, the free 3' end of the complementary strand anneals to an upstream segment of the same strand that includes the polymorphic site, such that the polymorphic site participates in the stem of the loop (see figures 14, 15). If the polymorphic nucleotide is complementary to the nucleotide near the 3' end of the strand a tight stem will be formed. If not, then a lower affinity interaction will exist and, at appropriately selected conditions, the stem will not form. Since the formation of the stem makes the 3' end of the strand no longer available for binding free primer, the amplification of the strand in which a perfect stem is formed is inhibited, as shown in Example 1. The length of the 5' extension on the primer can be varied, depending on the desired size of the loop, or on whether it is desirable to form alternative structures or enzyme recognition sites.



Some of the alternative structures that would be useful include: (i) recognition sites for various DNA modifying enzymes such as restriction endonucleases, (ii) a cruciform DNA structure, that could be very stable, or could be recognized by enzymes such as bacteriophage resolvases (e.g., T4E7, T7E1), or (iii) recognition sites for DNA binding proteins (preferably from thermophilic organisms) such as zinc finger proteins, catalytically inactive endonucleases, or transcription factors. The purpose of inducing such structures in an allele specific manner would be to effect allele specific binding to, or modification of, DNA. For example, consider a duplex formed only (or preferentially) by a strand from one allele that contains the recognition sequence for a thermostable restriction enzyme such as Taq I. Allele specific strand cleavage could be achieved by inclusion of (thermostable) Taq I during the PCR, resulting in complete inactivation of each cleaved template molecule and thereby leading to allele selective amplification. What are the limits of such an approach? One requirement is that there are no Taq I sites elsewhere in the PCR amplicon; another is that one of the two alleles must form a Taq I recognition sequence. The former limitation – which would limit the length of amplicons that could be allele specifically modified if only frequently occurring restriction sites were used – as well as the latter limitation – which requires that polymorphisms occur in restriction endonuclease recognition sites - can be addressed in part by designing a 5' primer extension, along with an internal primer loop, so that the recognition sequence for a rare cutting restriction endonuclease that (i) is an interrupted pallindrome, or (ii) cleaves at some distance from its recognition sequence is formed by the internal loop, while (i) the other end of the interrupted pallindrome, or (ii) the cleavage site for the restriction enzyme, occurs at the polymorphic nucleotide, and is therefore sensitive to whether there is a duplex or a (partially or completely) single stranded region at the polymorphic site. This scheme is illustrated in Figure 20. Preferred enzymes for PCR implementation of these schemes would include enzymes from thermophiles, such as Bsl I (CCNNNNN/NNGG) and Mwo I (GCNNNNN/NNGC).

Other alternative schemes would entail placing the stem-forming nucleotides internally, rather than at the end of the primer.

The experiments described above and in Example 1 are directed to stem formation during PCR, which requires that the stem be stable at an annealing temperature of ~50°C or greater.

However, isothermal amplification methods, such as 3SR and others, can also be used to achieve allele specific amplification. For isothermal amplification methods the loop forming sequences would likely be designed differently, to achieve maximum allele discrimination in secondary structure formation at 37°C, 42°C or other temperatures suited to amplification. This can be achieved by shortening the length of duplex regions. Example 1 gives typical lengths of duplex regions for PCR-based methods. Shorter duplex lengths can be tested empirically for isothermal amplification methods.

Our method does not address the intrinsic limitations of all PCR-based methods, but does provide superior performance compared to other procedures in the literature, excellent allele specificity can be achieved at fragment lengths of up to 4 kb.

## 5 7. Restriction endonuclease-based haplotyping methods

10 The first type of polymorphisms used to produce high density human genetic maps were restriction fragment length polymorphisms (RFLPs). RFLPs are polymorphisms, usually but not necessarily SNPs, that affect restriction endonuclease recognition sites. Initially RFLPs were identified, and subsequently typed, using Southern blots of genomic DNA. An RFLP was detected when the pattern of hybridizing species in a Southern blot (hybridized with a single copy probe) varied from sample to sample (i.e. from lane to lane of the Southern blot). Generally one detectable fragment would be identified in some lanes, one or two smaller fragments in other lanes, and both the large and smaller fragments in still other lanes, corresponding to homozygotes for the allele lacking the restriction site, homozygotes for the allele containing the restriction site and heterozygotes for the two alleles. The size difference between the restriction fragments lacking the polymorphic restriction site and those with the restriction site depends on the distance from the polymorphic restriction site to flanking, non-polymorphic sites for the same restriction enzyme.

20 In the past the location of polymorphic restriction sites and the sizes of the restriction products have generally been determined empirically. Although many restriction site polymorphisms have been converted to PCR assays by designing oligonucleotide primers flanking the polymorphic site these assays lack the character of the initial RFLP assays in which the restriction enzyme did all the work, and the size of the restriction fragments varied over a wide range.

25 In one aspect of this invention, RFLPs can be used to produce long range haplotypes, over distances of at least 5 kb, frequently over 10 kb and in some instances, using rarely occurring restriction sites, distances of up to 100 kb or greater. The basic approach is to: (i) select a DNA segment to be haplotyped (the exact boundaries will be constrained by the next step); (ii) identify a polymorphism, either within the segment, or, preferably, in flanking DNA, that alters a restriction enzyme recognition site for a restriction endonuclease (RE1). The outer bounds of the segment to be haplotyped are defined by the nearest occurrence of RE1 on either side of the polymorphic site.; (iii) Prepare genomic DNA from samples that are heterozygous for the polymorphism identified in step ii. It is desirable that the average length of the genomic DNA be greater than the length of the DNA fragment being haplotyped. (iv) Restrict the genomic DNA with the enzyme that recognizes the selected polymorphic site; (v) separate the restricted DNA

using any DNA size fractionating method suitable to the size range of the restriction fragments of interest (including gel electrophoresis; centrifugation through a gradient of salt, sucrose or other material, including use of step gradients; chromatography using sephadex or other material or other methods known in the art); (vi) isolate a first DNA fraction containing the larger restriction fragment and, optionally, a second DNA fraction containing the smaller restriction fragment and, if necessary, purify DNA from each fraction for PCR. It is not necessary that the fragments be highly enriched in the fractions, only that each of the one or more DNA fractions contain a significantly greater quantity of one allele than of the other. A minimum differential allele enrichment that would be useful for haplotyping is 2:1, more preferably at least 5:1 and most preferably 10:1 or greater. (vii) Genotype the polymorphic sites of interest in either one of the fractions (the one enriched for the larger allele or the one enriched for the smaller allele), or, optionally, determine genotypes separately in both size fractions. Since each fraction contains principally one allele, the genotype of the fractions provides the haplotypes of the enriched alleles. If only one fraction is genotyped, providing one haplotype, then the other haplotype can be inferred by subtracting the determined haplotype from the genotype of the total genomic DNA of the samples of interest. In a haplotyping project it is desirable to determine the genotypes in total genomic DNA of all samples of interest in advance of the haplotyping project, in order to determine, first, which samples actually require haplotype analysis (because they contain two or more sites of heterozygosity in the segment of interest), second, which samples are heterozygotes at the restriction site polymorphism selected for separation of the alleles by size, and are therefore suitable for analysis by the above method; third, the genotype of the total sample constrains the possible haplotypes, and provides a check on the accuracy of the haplotypes. Preferably the haplotype of both alleles are determined separately and compared to the genotype of the unfractionated sample. Samples that are not suitable for haplotype analysis with one restriction enzyme (because they are not heterozygous at the restriction site) can be analyzed with a different restriction enzyme, using the steps described above.

In another aspect, two restriction enzymes plus an exonuclease can be used in a haplotyping scheme that does not require a size separation step. In this method, illustrated in Figures 18, 19, 20, the initial steps are as above: (i) select a DNA segment to be haplotyped (the exact boundaries will be constrained by the next two steps); (ii) identify a polymorphism, either within the segment, or, preferably, in flanking DNA, that alters a restriction enzyme recognition site for a restriction endonuclease (RE1), the outer bounds of the segment to be haplotyped are defined by the nearest occurrence of RE1 on either side of the polymorphic site; (iii) identify a second restriction endonuclease (RE2) which cleaves only once within the segment to be haplotyped; (iv) prepare genomic DNA from samples that are heterozygous for the polymorphism identified in step ii, it is desirable that the average length of the genomic DNA be

greater than the length of the DNA fragment being haplotyped; (v) restrict the genomic DNA with RE1; (vi) block the ends of all cleavage products from exonuclease digestion (either by selecting an RE1 that produces termini not susceptible to exonuclease digestion – for example 3' protruding termini are resistant to cleavage by *E. coli* Exonuclease III; or by filling in recessed termini with nuclease-resistant modified nucleotides such as dideoxynucleotides, 5' amino-deoxynucleotide analogs, 2'-O-methyl nucleotide analogs, 2'-methoxy-ethoxy nucleotide analogs, 4-hydroxy-N-acetylprolinol nucleotide analogues or other chemically modified nucleotides (as described in PCT application, PCT US 99/22988, entitled METHODS FOR ANALYZING POLYNUCLEOTIDES); or by ligating adapters with nuclease resistant changes to the restriction termini); (vii) restrict with RE2. At this point, the two alleles in the DNA region of interest are in a different state. Allele A was cleaved in two by RE1 at the polymorphic site, both fragments were blocked from endonuclease digestion, and then RE2 cleaved *one* of the two fragments in two pieces, both of which have one end unprotected from exonuclease (a requirement of RE2 is that it produce termini that are susceptible to exonuclease digestion). Crucially, the fragment *not cleaved* by RE2, is still protected at both termini. Conversely, Allele B, lacking an RE1 site at the polymorphic site, was in one piece after RE1 digestion. RE2 digestion cleaved that one piece in two, both of which are susceptible to nuclease digestion, the consequence of which is the exonuclease digestion of both halves of the fragment (from the unprotected ends). Thus nuclease acts on the entire segment to be haplotyped in Allele B. (viii) After nuclease digestion, or at the same time, a small amount of a single strand specific nuclease may be added in order to destroy any single stranded regions left after the exonuclease treatment. This is important only if the first nuclease has no single strand nuclease activity (as is the case, for example, with *E. coli* Exonuclease III). Nuclease(s) can be inactivated, for example by heating, if necessary. (ix) A genotyping procedure can be used to determine the status of all polymorphic sites in the segment of Allele A that did not contain the site for RE2, and thus remained blocked at both ends during the exonuclease treatment. Since there is no (or little) Allele B remaining in the test tube, only the nucleotides corresponding to Allele A will be registered by the genotyping procedure, and they constitute the haplotype. A variety of nucleases can be used for this method, as well as combinations of nucleases, with, for example, one converting fragments with unprotected ends into single stranded DNA molecules and the other digesting single stranded DNA exo- or endonucleolytically. Specific nucleases useful for this method include *E. coli* Exonucleases I and III, Nuclease Bal-31 (which must be used with a suitable end protection procedure at step vi), as well as the single strand specific Mung Bean Nuclease, human cytosolic 3'-to-5' exonuclease and many other prokaryotic and eukaryotic exonucleases with processivity. Since large segments are more attractive as haplotyping targets than short ones the processivity of the nuclease may be a limit the utility of the method.

Therefore highly processive nucleases are preferred. Such nucleases may be either natural or modified by mutagenesis. As with other haplotyping methods, a minimum differential allele enrichment that would be useful is 2:1, more preferably at least 5:1 and most preferably 10:1 or greater. It is also preferable to haplotype the polymorphic sites of interest on both alleles in separate reactions. Alternatively, if the haplotype of only one allele is determined directly, then the other haplotype can be inferred by subtracting the known haplotype from the genotype of the total genomic DNA of the samples of interest. Haplotypes can be extended over long regions by the combined use of several restriction fragment length polymorphisms suitable for the method as outlined above.

In the future, with a complete sequence of many genomes, including the human genome, available, and hundreds of thousands, if not millions, of polymorphic sites identified it will be possible to design RFLP-based assays for the methods described above *in silico*. That is, one will be able to identify, for any DNA segment of interest, the flanking restriction sites for any available restriction enzyme, and the subset of those sites that are polymorphic in the human (or other) population. Using criteria such as desired fragment location, desired fragment length, desired difference in length between two alleles (for separation by size) or location of a suitable site for R2 (for exonuclease removal of one allele) (for allele enrichment by selective exonuclease digestion), it will be possible to automate the design of RFLP assays. In another aspect of this invention a program for automatically designing experimental conditions, including restriction endonucleases and either electrophoretic (or other) separation conditions, or exonucleases, given the constraints just described can be executed.

In another aspect, a polymorphic restriction endonuclease site can be exploited for haplotyping in conjunction with an amplification step, wherein first a genomic DNA sample is treated with a restriction endonuclease that cleaves at a polymorphic site, and second an amplification is performed spanning the restriction cleavage site. If one of the two alleles present in a DNA sample contains the recognition site for the enzyme then it will not be amplified due to strand scission of all template molecules. The other allele will be amplified, and the amplification product can subsequently be diluted and subjected to a genotyping procedure. The set of genotypes obtained constitute the haplotype of the allele lacking the polymorphic RE1 site. In this method, illustrated in Figures 25 and 26, the initial steps are as above: (i) select a DNA segment to be haplotyped (the exact boundaries will be constrained by the next step); (ii) identify a polymorphism, either within the segment, or, preferably, in flanking DNA, that alters a restriction enzyme recognition site for a restriction endonuclease (RE1). The outer bounds of the segment to be haplotyped are defined by the nearest occurrence of RE1 on either side of the polymorphic site. (iii) Prepare genomic DNA from samples that are heterozygous for the polymorphism identified in step ii. It is desirable that the average length of the genomic DNA be

greater than the length of the DNA fragment being haplotyped. (iv) Restrict the genomic DNA with RE1; (v) perform an amplification, for example a PCR amplification, using forward and reverse primers located on opposite sides of the polymorphic RE1 site, but within the DNA segment subtended by the flanking, non-polymorphic, RE1 sites. (vi) dilute the amplified DNA (optional; useful mainly if the genotyping procedure of the next step is amplification-based); (vii) subject the amplified DNA to genotyping tests for one or more polymorphisms that lie within the amplified segment. Virtually any genotyping method will work. One preferred genotyping method (not requiring the dilution of step vi) is primer extension, followed by electrophoretic or mass spectrometric analysis. Primers are positioned just upstream of one or more polymorphic sites in the amplified segment, extended in an allele specific manner and analyzed using methods known in the art. This method can also be used in conjunction with allele specific priming experiments of this invention, in order to boost specificity of allele amplification.

Restriction endonuclease sites that flank the target segment can be exploited to produce optimally sized molecules for allele selection. For example, a heterozygous DNA sample can be restricted so as to produce two allelic DNA fragments that differ in length, and consequently differ from one another by the presence or absence of a binding site for an allele specific binding reagent. Because of the ease of restriction endonuclease digestion, and the possibility of cleaving just outside the target DNA segment to be haplotyped (thereby producing the maximal size DNA fragment that differs in respect to the presence/absence of a single binding site), complete restriction is a preferred method for controlling the size of DNA segments prior to allele enrichment.

## 8. Imaging-based haplotyping methods

### (a) Optical mapping technology

One way to determine haplotypes would be to use microscopy to directly visualize a double stranded DNA molecule in which the sequence of the DNA at polymorphic sites is revealed visually. David Schwartz and colleagues have developed a family of methods for the analysis of large DNA fragments on modified glass surfaces, which they refer to as optical mapping. Specifically, Schwartz and colleagues have devised methods for preparing large DNA fragments, fixing them to modified glass surfaces in an elongated state while preserving their accessibility to enzymes, visualizing them microscopically after staining, and collecting and processing images of the DNA molecules to produce DNA restriction maps of large molecules. (Lai Z, Jing J, Aston C, et al. A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nat Genet.* 1999 Nov;23(3):309-13; Aston C, Mishra B, Schwartz DC. Optical mapping and its potential for large-scale sequencing projects. *Trends Biotechnol.* 1999 Jul;17(7):297-302; Aston C, Hiort C, Schwartz DC. Optical mapping: an approach for fine

mapping. *Methods Enzymol.* 1999;303:55-73; Jing J, Reed J, Huang J, et al. Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *Proc Natl Acad Sci U S A.* 1998 Jul 7;95(14):8046-51.) Many of the imaging and image analysis steps have been automated. (see articles cited above and: Anantharaman T, Mishra B, Schwartz D. Genomics via optical mapping. III: Contigging genomic DNA. *Ismb.* 1999;(6):18-27.) Many of the optical mapping methods have also been described in United States Patent 5,720,928.

The optical mapping methods of Schwartz and colleagues have so far been largely confined to the generation of restriction endonuclease maps of large DNA segments or even genomes by treating immobilized, surface bound double stranded DNA molecules with restriction endonucleases, and to a lesser extent, studies of DNA polymerase on single DNA molecules. For example, a complete BamH I and Nhe I restriction map of the genome of *Plasmodium Falciparum* has been made using optical mapping. The average fragment length of analyzed fragments was 588 – 666 kb, and the average coverage of the map was 23 X for Nhe I and 31 X for BamH I. (That is, on average, each nucleotide of the genome was present in 23 or 31 different analyzed fragments. This high level of redundancy provides higher map accuracy.) *P. falciparum* has a genome length of ~24.6 megabases, so, taking into account the 31 X redundancy of the BamH I map, ~763 mb were analyzed. The human genome, at ~3,300 mb, is only about 4 times larger than the scale of this experiment (albeit at 1X coverage, which would be insufficient for highly accurate results). However, it should be possible, using a higher density of DNA fragments, and/or a larger surface, to prepare glass slides with fragments corresponding to several equivalents of the human genome. Statistically reliable haplotyping results would be obtainable from such DNA preparations, using the methods described below. As an alternative to whole genome preparations, size selected fractions of the genome, or long range amplification products could also be used for the haplotyping methods described below.

It is an aspect of the present invention that optical mapping and related methods can be exploited for determination of haplotypes. Several methods can be coupled with optimal mapping technology to determine haplotypes: (i) Restriction endonuclease digestion using enzymes that cleave at polymorphic sites on the DNA segment to be haplotyped, (ii) addition of PNAs corresponding to polymorphic sites to form allele specific D-loops, (iii) addition of sequence specific DNA binding proteins that recognize sequences that are polymorphic, and that consequently bind only to one set of alleles. The various types of allele specific DNA binding proteins described above are all useful in this aspect, however, the versatility in terms of sequence recognition and high affinity binding of zinc finger proteins make them a preferred class of DNA binding proteins. (iv) addition of other sequence specific binding molecules described in this application.

A haplotyping method based on zinc fingers and optical mapping would consist of the following steps: (i) prepare fixed, elongated DNA molecules according to the methods of Schwartz, (ii) add zinc fingers that recognize polymorphisms in a DNA segment to be haplotyped. Preferably the zinc fingers are synthesized with a detectable label, for example by making a fusion protein, or alternatively they are post-translationally labelled. Preferably different zinc fingers are labelled (whether by making fusion proteins or by post-translational chemical modification) with two or more different methods that result in detectable differences. Ideally at least two different labels are used for the zinc finger proteins. The reason that more than one type of signal is preferable is that when two or more zinc finger proteins are bound to a DNA molecule there will be a pattern to the detectable labels. The pattern, as well as the distance between the zinc finger proteins, provides a signature that helps identify the DNA molecule to which the proteins are bound.

#### (b) Atomic force microscopy

In another aspect of the invention, atomic force microscopy can be used in a manner substantially similar to that described above for optical mapping. That is, detectable structures can be formed at polymorphic sites by addition of DNA binding proteins, preferably zinc finger proteins, or by forming other detectable complexes at polymorphic sites. Another method for forming detectable structures at polymorphic sites is strand invasion, preferably using PNA molecules. By appropriate design and optimization of PNA molecules an allele specific strand invasion can be effected.

As with the haplotyping methods based on optical mapping, the haplotyped molecules may be either PCR products or genomic DNA fragments.

### C. ApoE Genotypes and Haplotypes

The Apolipoprotein E gene (ApoE) encodes a well studied protein (APOE) central to lipoprotein metabolism. The existence of three major allelic forms of ApoE (referred to as e2, e3 and e4) has been known for over 2 decades. The well established three allele classification of ApoE is based on two polymorphisms in the coding sequence of the ApoE gene, both of which result in cysteine vs. arginine amino acid polymorphisms in APOE protein at positions 112 and 158 of the mature protein. DNA-based diagnostic tests for ApoE have been available since the 1980s.

The ApoE e4 allele has been consistently correlated with elevated total cholesterol, elevated LDL cholesterol, low levels of ApoE protein and increased risk of coronary heart disease (CHD). The CHD risk attributable to e4 is apparent even after correcting for cholesterol levels and other CHD risk factors (smoking obesity, diabetes, blood pressure). The e4 allele is



also a risk factor for late onset Alzheimer's disease (AD), apparently due to effects on the rate of disease progression. Presence of the ApoE e4 allele also portends a poor prognosis for patients with a variety of other neurological diseases (stroke, brain trauma, amyotrophic lateral sclerosis and other diseases) and psychiatric diseases (e.g. schizophrenia), compared to patients without an e4 allele.

In addition to effects on disease risk and disease prognosis there are reports that ApoE genotype predicts response of AD patients to medications. In particular, the response of Alzheimer's disease patients to acetylcholinesterase inhibitors has been studied by several groups. ApoE genotype may also be useful for predicting patient response to other medical treatments, particularly treatments for neurological and cardiovascular diseases.

However, despite the many genetic associations described above, diagnostic tests for determining ApoE genotype are not widely used, nor is ApoE genotyping widely used for prognostic or pharmacogenetic testing. To the contrary, a large number of studies address the limitations of ApoE as a diagnostic marker, particularly in the setting of AD diagnosis. The conclusion of most of these studies is that testing for the e2, e3 and e4 alleles does not provide a sufficiently sensitive or selective test to justify use outside of clinical research. Concern has also been expressed that, because in many settings ApoE testing results do not affect medical decision making, there is little reason to obtain information on ApoE genotype.

Recent studies of the ApoE gene in a number of laboratories have led to identification of several new DNA polymorphisms. The biological effects and medical import of these new polymorphisms has not been established, although some studies suggest that polymorphisms in the promoter affect ApoE transcription rates. Most published work has been limited to the analysis of individual polymorphisms or sets of only a few polymorphisms and their effect on one or two biological or clinical endpoints.

In the present application we describe multiple previously unreported, polymorphisms in the ApoE gene. Also described are methods for determining the ApoE genotype and haplotype of unknown samples, e.g., using methods of this invention. These new genotyping and haplotyping methods will enable more accurate measurement of the contribution of variation in the entire ApoE gene (promoter, exons, introns and flanking DNA) to variation in serum cholesterol, CHD risk, AD risk, prognosis of patients with neurodegenerative diseases or brain trauma, responses of patients to various treatments and other medically important variables described herein. These improved ApoE tests may provide the degree of sensitivity and selectivity required for successful development of diagnostic, prognostic or pharmacogenetic tests for neurological, psychiatric or cardiovascular disease, either alone or in combination with genetic tests for other relevant genes.

Apolipoproteins are found on the surface of various classes of lipoproteins – membrane bound particles which transport lipids (mainly cholesterol and triglycerides) throughout the body, including the brain. The function of apolipoproteins is to direct lipoproteins to specific cells that require lipids, for example cells that store fat. The apolipoproteins bind to specific receptors on the surface of lipid requiring cells, thereby directing the transport of lipids to the target cell. Apolipoprotein E is one of about a dozen apolipoproteins on blood lipoproteins, but it is the major apolipoprotein in the brain.

One important function of ApoE in the brain is to transport lipids to cells that are performing membrane synthesis, which often occurs as a response to acute or chronic brain injury. After injury there is usually extensive synaptic remodeling as the surviving neurons receive new inputs from cells that were formerly wired to injured cells. This neuronal remodeling, or plasticity, is an important part of the physiologic response to the disease process and modulates the course of disease. Patients with low ApoE levels or impaired ApoE function have impaired neuronal plasticity. In Alzheimer's disease one injured brain region is the cholinergic pathways of the basal forebrain and elsewhere. The degree of neuronal remodeling in such areas may affect the response to cholinomimetic therapy. Thus impaired brain lipid transport alters patterns of neuronal remodeling in cholinergic (and other) pathways and thereby potentially affects response to acetylcholinesterase inhibitors and possibly other cholinergic agonists.

The ApoE4 allele is a major risk for Alzheimer's disease, perhaps because it is expressed in brain at lower levels than the E2 or E3 alleles, and thus impairs neuronal remodeling. The E2 allele is mildly protective for AD. Several clinical trials for Alzheimer's Disease drugs, including both acetylcholinesterase inhibitors and vasopressinergic agonists, have shown significant interactions with ApoE genotype and sex. The E4 allele has been associated with lack of response to acetylcholinesterases.

The relative risk of AD conferred by the E4 allele varies almost ten-fold between different populations. The highest relative risk (RR) has consistently been reported in the Japanese, who have a 30-fold RR in E4/E4 homozygotes relative to E3/E3 homozygotes. African and Hispanic E4/E4 homozygotes have relative risks of only ~3-4-fold. On the other hand, in the presence of an E4 allele the cumulative risk of AD to age 90 is similar in all three groups (Japanese, Hispanics and Africans). This suggests that other factors contribute significantly to the causation of AD in the non-Japanese populations. It may be that these non-E4 AD patients are the best responders to acetylcholinesterases. If true, this may account for a lack of response in Japanese, where the fraction of patients with ApoE4 mediated AD appears to be the highest in the world.

It is well established that the three common alleles at the ApoE locus are correlated with risk of AD in various populations. Recent studies have also shown that ApoE genotype correlates with response of AD patients to two classes of drugs. Specifically, Poirier et al. demonstrated an interaction of apoE genotype, sex and response of AD patients to the cholinomimetic drug tacrine, while Richard, et al. showed an interaction between apoE genotype and response to an investigational noradrenergic/vasopressinergic agent, S12024. In both studies the analysis was restricted to analysis of the two amino acid variances that determine the three common ApoE alleles. Other variances have been described at the ApoE locus, including promoter variances, that may affect ApoE function. Also, studies have been published associating polymorphisms in other genes with risk of late onset AD. However, there have been no investigations of the effect of variation at these loci on response to cholinomimetic drugs.

There are two FDA approved drugs for therapy of Alzheimer's Disease (tacrine, donepezil), and at least a dozen additional agents in late stage clinical trials or under FDA review. The FDA approved drugs work by inhibiting acetylcholinesterase, thereby boosting brain acetylcholine levels. This symptomatic therapy provides modest benefit to less than half of treated patients but does not affect disease progression. Available evidence suggests the products in the pipeline, which likewise partially reverse symptoms without affecting the underlying disease process, will also be of modest benefit to some patients. Despite their limited efficacy, these drugs will be expensive. They will also likely be associated with serious adverse effects in some patients. As a result, the cost of providing a modest benefit to a limited number of Alzheimer's Disease patients will be high.

As more AD therapeutics becomes available, physicians will face the difficult task of differentiating between multiple products. These products may produce similar response rates in a population, however, the crucial decision clinicians face is selecting the appropriate therapeutic for each individual AD patient at the time of diagnosis. This is particularly the case if there are several therapeutic choices, while only one of which may be optimal for a particular patient. This selection is critical because failure to provide optimal treatment at the time of diagnosis may result in a diminished level of function during a period when the greatest benefit could be achieved. Inadequate treatment may continue for some time because measures of clinical response in AD are notoriously imprecise; six months or longer may pass before it is clear whether a drug is working to a significant degree. During this time, the disease continues to progress which may limit the efficacy of a second drug or therapeutic regimen. A test that could predict likely responders to one or more AD drugs would thus be of great value in optimizing patient care and reducing the cost of ineffective treatment.

Data has been published suggesting that ApoE genotype may be such a test. Specifically, Farlow, Poirier and colleagues have shown that female patients with the APOE ε4 allele do not

respond to tacrine, while female patients with the  $\epsilon 2$  and  $\epsilon 3$  alleles have significant response; males do not respond significantly regardless of genotype. Conversely, Richard et al. have demonstrated that patients with the  $\epsilon 4$  allele, but not the  $\epsilon 2$  and  $\epsilon 3$  alleles, have a statistically significant response to S12024, an enhancer of vasopressinergic/noradrenergic signalling. Thus the two drugs – one an acetylcholinesterase inhibitor and the other a vasopressinergic/noradrenergic agonist – are useful in different groups of patients, delimited by ApoE genotype.

The ability to predict response to therapy for progressive debilitating diseases like AD would be of enormous clinical importance as there is generally only one opportunity to treat patients with these diseases at their maximal level of functioning; any delay in selecting optimal therapy represents a lost opportunity to preserve the maximal possible level of function. With multiple drugs in development for AD it will become increasingly important to predict the best drug for each patient.

#### *Screening the ApoE gene for variation*

In order to better understand genetically encoded functional variation in the ApoE gene and its encoded product we systematically cataloged genetic variation at the ApoE locus. The ApoE genomic sequence is represented in GenBank accession AB012576. The gene is composed of four exons and three introns. The transcription start site (beginning of first exon) is at nucleotide (nt) 18,371 of GenBank accession AB012576, while the end of the transcribed region (end of the 3' untranslated region, less polyA tract) is at nt 21958 (Table 2).

We designed PCR primer pairs to cover the ApoE genomic sequence from nucleotides 16,382 – 23,984. Thus, our analysis began 1,989 nucleotides upstream of the transcription start site, extended across the entire gene and ended 2,026 nucleotides after the final exon. This segment of DNA was chosen to allow us to uncover any polymorphisms that might affect upstream, downstream or intragenic transcriptional regulatory sequences, or that could alter transcribed sequences so as to affect RNA processing (splicing, capping, polyadenylation), mRNA export, translation efficiency, mRNA half life, or interactions with mRNA regulatory factors, or that could affect amino acid coding sequences.

Separately, the ApoE cDNA was screened for polymorphism. The ApoE cDNA sequence was obtained from GenBank accession K00396, which covers 1156 nt. Nucleotides 43 through 1129 were screened by DNA sequencing.

We also searched for polymorphisms in a putative ApoE enhancer element located ~15 kb 3' of the end of the ApoE gene, in the expectation that polymorphisms in a regulatory element might affect ApoE levels. The enhancer sequence is in the same GenBank accession as the ApoE gene (AB012576). The segment screened for polymorphism extends from nt 36,737 to 37,498.

Exemplary polymorphism screening methods are described in Example 3. Briefly a panel of 32 subjects of varying geographic, racial and ethnic background were selected for screening.

A total of 20 polymorphic sites were identified, several of which correspond to polymorphisms previously reported in the literature (see Table 2). We also report unique haplotypes that have been observed with these polymorphisms. Table 3 shows an analysis of the haplotypes present in a subset of nine polymorphic sites. These haplotypes were determined using the methods described in detail in Example 1.

Table 4 provides the sequence of 42 additional haplotypes of the ApoE gene. In any given haplotype, the ApoE sequence between the listed nucleotides (e.g. between 16,541 and 16,747) is generally identical to that in the GenBank AB012576, however there may be additional polymorphic sites not listed in this table. Such additional variant sites do not lessen the utility of the haplotypes provided. Where no sequence is provided at a particular site in a particular haplotype (e.g. position 18145 of haplotype 4) it is understood that either of the two nucleotides that appear elsewhere in the column (T or G under column 18145) could appear at the indicated site.

Other haplotypes of the ApoE gene are shown in Table 5. In this table a useful group of haplotypes is shown. These haplotypes are specified by SNPs at positions 16747, 17030, 17785, 19311, and 23707 (as shown in rows 1-4 of the table) or by SNPs at a subset of these positions: 17785, 19311, and 23707 (rows 5-8); 17030, 19311, and 23707 (rows 9-12); 16747, 19311, and 23707 (rows 13-16); 17030, 17785, and 23707 (rows 17-20); 16747, 17030, 19311, and 23707 (rows 21-24); or 16747, 17785, 19311, and 23707 (25-28 of the table). One useful aspect of these haplotypes is that they closely parallel the classic phenotypes as indicated in the column on the far right. That is, the haplotype GCAGC in row 1 identifies the alleles designated E3 by the classic ApoE test; and GCAGA, in row 3, specify the alleles designated E4 by the classic ApoE test; and GCAGA, in row 4, identifies the alleles designated E2 by the classic ApoE test. The haplotypes in rows 5-28 are simpler versions of those in rows 1-4, with the corresponding classic ApoE genotype/phenotypes indicated in the GENOTYPE column. It should be noted that the polymorphisms that specify the classic ApoE alleles are encoded by nucleotides 21250 (first position of codon 112 of the mature ApoE protein) and 21388 (first position of codon 158) of the mature ApoE protein). Nucleotides 21250 and 21388 are not elements of the haplotypes specified in Table 4. In other words, the haplotypes in Table 4 are based upon SNPs that are completely different from the SNPs that form the basis of current ApoE allele classifications and genotype/haplotype tests. Thus, determining a haplotype or pair of haplotypes in a sample by a method that comprises examining any of the combinations of SNPs provided in Table 4, below constitutes a novel method for determining the classic ApoE genotype/phenotype status of a sample.

Preferably, a haplotype or haplotypes specified in the Table 5 are determined in conjunction with at least one additional ApoE SNP specified herein (see Table 4). To constitute a new set of haplotypes.

5 Preferably, the at least one additional SNP (beyond those in Table 5) divides at least one of the three classical ApoE phenotypes into two haplotype groups. For example, addition of the C/T polymorphism at nucleotide 21349 to the group in Table 5 divides the E3-like haplotypes into two groups; those with C at 21349 and those with T at 21349. Addition of the T/C polymorphism at nucleotide 17937 to those in Table 5 divides the E2-like haplotypes into two groups: those with a T at 17937 and those with a C at 17937. Such subgroups are more likely to  
10 correspond to biologically and clinically homogeneous populations than the classic  $\epsilon 2$ ,  $\epsilon 3$ ,  $\epsilon 4$  classification.

005207 " 32076960

## Examples

### Example 1. Haplotyping Method Using Hairpin Inducing Primers for Allele Specific PCR

A primer is designed which contains at least two different regions. The 3' portion of the primer corresponds to the template DNA to be amplified. The length of this region of the primer can vary but should be sufficient to impart the required specificity to result in amplification of only the region of cDNA or genomic DNA of interest. Additional nucleotides are added to the 5' end of the primer which are complementary to the region in the sequence which contains the nucleotide variance. Following two rounds of PCR, the added tail region of the primer is incorporated into the sequence. Incorporation of the added nucleotides causes the reverse strand complementary to the primer strand to form a hairpin loop if the correct nucleotide is present at the site of variance. The hairpin loop structure inhibits annealing of new primers and thus further amplification.

Primers with the above characteristics were designed for haplotyping of the dihydropyrimidine dehydrogenase (DPD) gene. The DPD gene has two sites of variance in the coding region at base 186 (T:C) and 597 (A:G) which result in amino acid changes of Cys:Arg and Met:Val, respectively (Figure 27). The second site at base 597 is a restriction fragment length polymorphism (RFLP) which cleaves with the enzyme BsrD I if the A allele is present. Primers were designed which would result in amplification of one or the other allele depending which base was present at the site of variance at base 186 (Figure 28). The bases in white correspond to the region of the primer which is complementary to the DPD sequence. The green base is the variant base in the target sequence. The bases in red and blue at bases added to the 5' end of the primer which should form a hairpin loop following incorporation into the PCR product. The blue base is the added base which hybridizes to the variant base and is responsible for the allele discrimination of the hairpin loop. The DPDNSF primer contains only the DPD complementary sequence and will not result in allele specific amplification. In Figure 29 shows hybridization of the non-specific DPDNSF primer to both the T and A allele of the DPD target sequence and the 5' end of the PCR product generated by amplification using this primer. Figures 30 and 31 are the corresponding diagrams as shown in Figure 29, for primers DPDASTF and DPDASCF. Notice that the added bases are incorporated into the PCR fragment following amplification. Figure 32 shows the most stable hairpin loop structures formed with the reverse strand of the PCR product made using the DPDNSF primer using the computer program Oligo4. Only the reverse strand is shown because this would be the strand to which the DPDNSF primer would hybridize on subsequent rounds of amplification. The hairpin loops are either not stable or have a low melting temperature. Figures 33 and 34 are the corresponding diagrams for the hairpin loops formed in the reverse strands of the PCR products generated using primers DPDASCF and DPDASTF, respectively. Amplification using primer DPDASCF of the T allele

results in the ability to form a very stable hairpin loop with a melting temperature of 83°C (Figure 33). In contrast, amplification of the C allele with primer DPDASCF generates a hairpin loop with a melting temperature of only 42°C. The converse is true for the primer DPDASTF. Amplification of the C allele of DPD results in the formation of a very stable hairpin loop (100°C) while amplification of the T allele results in the formation of a much less stable hairpin (42°C) (Figure 34).

Figures 35, 36 and 37 depict the primer hybridization and amplification events when further amplification is attempted on the generated PCR fragments. The DPDNSF primer is able to effectively compete with the hairpin structures formed with both the T and C allele of the DPD gene and thus amplification of both alleles proceeds efficiently (Figure 35). The DPDASCF primer (Figure 36) is able to compete for hybridization with the hairpin loop formed with the C allele because its melting temperature is higher than the hairpin loop's (60°C compared to 42°C). The hairpin loop formed on the T allele however, has a higher melting temperature than the primer and thus effectively competes with the primer for hybridization. The hairpin loop inhibits PCR amplification of the T allele which results in allele specific amplification of the C allele. The reverse is true for the primer DPDASTF (Figure 37). The hairpin loop structure has a higher melting temperature than the primer for the C allele and a lower melting temperature than the primer for the T allele. This causes inhibition of primer hybridization and elongation on the C allele and results in allele specific amplification of the T allele.

The ability to use this for haplotyping is diagrammed in Figure 28 using a cDNA sample whose haplotype is known to be : Allele 1 – T<sup>186</sup>:A<sup>597</sup>, Allele 2 – C<sup>186</sup>:G<sup>597</sup>. The size of the fragments generated by a BsrD I from a 597 bp generated by amplification with the primers DPDNSF, DPDASTF, and DPDASCF, depend on whether the base at site 597 is an A or a G. Restriction digestion by BsrD I is indicative of the A base being at site 597. If a fragment has the A base at 597, three fragments will be generated of lengths 138, 164 and 267 bp. If the G base is at site 597 only two fragments will be generated of lengths 164 and 405 bp. If a sample is heterozygous for A and G at site 597, you will generate all four bands of 138, 164 (2x), 267 and 405 bp. The expected fragments generated by BsrD I restriction for each of the primers is indicated in the box in Figure 38.

Figure 39 shows a picture of an agarose gel run in which each of the primers was used to amplify the cDNA sample heterozygous at both sites 186 and 597 followed by BsrD I restriction. The DPDNSF lane shows the restriction fragment pattern for the selected cDNA using the DPDNSF primer indicating that this sample is indeed heterozygous at site 597. However, using the same cDNA sample and the primer DPDASTF (DPDASTF lane), the restriction pattern correlates to the pattern representative of a sample which is homozygous for A at site 597.



Because the DPDASTF primer allows amplification of only the T allele, the haplotype for that in the sample must be T<sup>186</sup>:A<sup>597</sup>. The restriction digest pattern using the primer DPDASCF (DPDASCF lane) correlates with the expected pattern for there being G at site 597.

Amplification of the cDNA sample with the primer DPDASCF results in amplification of only the C allele in the sample. Thus the haplotype for this allele must be C<sup>186</sup>:G<sup>597</sup>. This demonstrates that primers can be designed that will incorporate a sequence into a PCR product which is capable of forming a hairpin loop structure that will inhibit PCR amplification for one allele but not the other allele even if there is only a single base pair difference between the two alleles. This can be exploited for allele specific amplification and thus haplotyping of DNA samples.

Alternatively, it may also be possible to form a hairpin structure at the 5' end of the PCR product which is stable enough to keep the polymerase from extending through the region. This may be possible by incorporating into the primer modified nucleotides or structures that when they hybridized to the correct base they would form a structure stable enough to inhibit read through by a polymerase.

This invention is meant to cover any method in which a stable secondary structure is formed in one or both strands of a PCR product which inhibits further PCR amplification. The secondary structure is formed only when the correct base or bases are present at a known site of variance. The secondary structure is not formed when the incorrect base or bases are present in the PCR product at the site of variance allowing further amplification of that product. This allows the specific amplification of one of the two possible alleles in a sample specific allowing the haplotyping of that allele.

## **Example 2. Genotyping of an ApoE variance by mass spectrometry analysis of restriction enzyme generated fragments**

The following example describes the genotyping of the variance at genomic site 21250 in the ApoE gene which is a T:C variance resulting in a cysteine to arginine amino acid change in amino acid 176 in the protein. Two primers were designed to both amplify the target region of the ApoE gene and to introduce two restriction enzyme sites (Fok I, Fsp I) into the amplicon adjacent to the site of variance. Figure 40 shows the sequence of the primers and the target DNA. The Apo21250-LFR primer is the loop primer which contains the restriction enzyme recognition sites and the ApoE21250-LR primer is the reverse primer used in the PCR amplification process. The polymorphic nucleotide is shown in italics. The following

components were mixed together in a 200 µl PCR tube for each genotyping reaction. All volumes are given in µl.

	A.	10x PCRx buffer (Gibco/BRL, cat# 11509-015)	2
5	B.	2 mM dNTP mix	2
	C.	50 mM MgSO <sub>4</sub>	0.8
	D.	PCR enhancer (Gibco/BRL, cat# 11509-015)	4
	E.	20 µM ApoE21250-LFR primer	1
	F.	20 µM ApoE21250-LR primer	1
10	G.	Patient genomic DNA 20 ng/ul	0.5
	H.	Platinum Taq DNA polymerase (Gibco/BRL, cat# 11509-015)	0.1
	I.	deionized water	8.6

The reactions were cycled through the following steps in MJ Research PTC 200 thermocyclers:

A.	94°C	1 min.	1 cycle
B.	94°C	15 sec.	B-D 45 cycles
C.	55°C	15 sec.	
D.	72°C	30 sec.	
E.	15°C	indefinitely	hold

The sequence of the amplicon for both the T allele and the C allele following amplification is shown in Figure 35. Five µl of each reaction were removed and analyzed by agarose gel electrophoresis to ensure the presence of sufficient PCR product of the correct size. The following components were mixed together for the restriction enzyme cleavage of the DNA.

25 Platinum Taq antibody (Taquench, Gibco/BRL cat# 10965-010) was added to inhibit any potential filling in of the 3' recessed end created by Fok I cleavage. All volumes are in µl.

	A.	10x New England Biolabs buffer #2	2
	B.	Fok I 4 units/µl (New England Biolabs, cat# 109S )	0.3
30	C.	Fsp I 5 units/µl (New England Biolabs, cat#135S )	0.2
	D.	Platinum Taq antibody (Gibco/BRL, cat# 11509-015)	0.2
	E.	PCR reaction	15
	F.	deionized water	2.4

35 The above reactions were incubated at 37°C for 1 hour. The cleavage sites for each amplicon are shown. Following incubation, the reactions were purified by solid phase extraction and eluted in

a volume of 100 µl of 70% acetonitrile water mix. The samples were dried in a Savant AES 2010 speed vac for 1 hour under vacuum and heat. The samples were resuspended in 3 µl matrix (65 mg/ml 3-hydroxy-picolinic acid, 40 mM ammonium citrate, 50% acetonitrile) and spotted on the Perseptive Biosystems 20x20 teflon coated plate. Samples were analyzed on the Perspective Biosystems Voyager-DE Biospectrometry™ Workstation.

### Example 3.

#### Screening the ApoE gene for polymorphism

PCR primers were selected automatically by a computer program that attempts to match forward and reverse primers in terms of GC content, melting temperature, and lack of base complementarity. The parameters of the program were set to select primers approximately 500 base pairs apart from each other, with at least 50 base pairs of overlap between adjacent PCR products. Primers were received in 96 well microtiter plates, resuspended in sterilized deionized water at a concentration of 5 pmoles/ul. PCR reactions were set up using a programmed Packard robot to pipet a master mix of 1X PCR buffer, polymerase and template into 96 well plates. Starting PCR conditions were: 10 mM Tris (pH 8.3), 50 mM KCl, 1.5 mM MgCl<sub>2</sub>, 0.2mM dNTPs, 0.83 uM forward and reverse primers, 0.7 Units of AmpliTaq Gold (PE Corp) and 25 ng of genomic template, in a volume of 30 ul. Cycling was done on MJ PTC200 PCR machines with the following cycle conditions: denature 12 minutes at 95°C followed by 35 cycles of: denature 15 seconds at 94°C, anneal 30 seconds at 60°C, extend 45 seconds at 72°C, followed by a ten minute extension at 72°C. PCR success was then tested by analyzing products on 6% Long Ranger acrylamide gels. Products passed if they exhibited clean bands stronger than a 15 ng standard, with little to no secondary amplification products. Efforts to optimize conditions for failed PCR products began with systematic variation of temperature, cosolvents (particularly PCR enhancer from GIBCO/BRL) and polymerase (Platinum Taq from GIBCO/BRL vs. AmpliTaq Gold). PCR products not optimized by these modifications were discarded and one or two new PCR primers were ordered and the process repeated until successful amplicons were produced.

Optimized PCR primer pairs were used to perform DNA cycle sequencing using ABI BigDye DNA sequencing kits according to instructions provided with the kits, except kit reagents were diluted 1:8 and A, G, C and T reactions were set up robotically in a volume of 20 ul.

Sequencing reactions were run on ABI 377 or ABI 3700 automated DNA sequencing instruments. ABI 377 and ABI 3700 run times were similar, approximately 4 hours at approximately 5000 volts. Data was collected automatically using ABI collection software. The quality of DNA sequencing reactions was assessed automatically and numerically scored using

the program PHRED. Only DNA sequence of quality level 30 or higher was considered acceptable for analysis.

Raw sequencing reactions were then imported into a custom database and analyzed using PHRED, PHRAP and POLYPHRED, and then the CONSED viewer was used to visually inspect the data and verify varainces. The custom database was used to track all samples in process and serve as a virtual notebook reference for all sample handling steps as well as data generation, manipulation and presentation

005201" 820/6960

	dA	dC	dG	dT	BrdU
dATP					
dCTP	24.0				
dGTP	16.0	40.0			
dTTP	9.0	15.0	25.0		
BrdUTP	55.8	79.8	39.8	64.8	

Table 1. Mass differences between the nucleotides dATP, dCTP, dGTP, dTTP, and BrdUTP.

Table 2**ApoE genomic sequence (GenBank accession AB012576) with polymorphisms indicated**

5 (partial sequence of the accession)

14701 ctggtggagc atctgatggg tgtttgggcc aagctggagc tttgtccatc ccctcttatt  
14761 tttctgcact tgactctctt atttttctga gactgggtct cctctgtcgc ccaggctaga  
14821 gtgcagcagt gcaactgcgg ctcaactgcag cctccacctc ccgggctcaa gcagccttcc  
10 14881 cacctcagcc tcctgagtag ctaggaccac aggtgtatgc caccaggccc agctaatttt  
14941 tttgatagtt ttgggagaca tgggggtttc accatgttgc ccaggctggg ctggaactcc  
15001 tggactcaag ccttggcctc ccaaagtgtc gggattatag gtgtgagcca ccacaccag  
15061 ccagggtaga aggcactttg gaagcctcga gcctgcccc ttcattctac gttagtggaa  
15121 actgaggctt ccagaggttt caaggtcaca actaaatcca gaacctcatc tcaggcacac  
15181 tggtcgtagt cccaatgtcc agtcttaagt cttcttggat atctgtggct cacagatttt  
15241 ggggtgttga gcctcctgct gagcactgct ggggccacag cggtgaccag ccctgtcttc  
15301 acgggactca gtgagaggaa cagattcatc cgcagagtgg gcaggactag gttgggggaa  
15361 cccaggggtc tagagggctt ttcagagggc aggggtcact gagcggagag cagaggagga  
15421 gtgagccatt tgctccagcg tgaagttgtt ggtgtgatgg ggtttcaggg tggcaggagc  
15481 agtgtgggta aaggtctgga agctgtcggc atgtggctgg tatccaaggt ggccaggaac  
15541 tctgcatgga tatggtggga agctggcacg cctctcacct cagctcttcc ctgcaggctc  
15601 tgtggatagc aactggatcg tgggtgccac gctggagaag aagctccac ccctgcccct  
15661 gacactggcc cttggggcct tcctgaatca ccgcaagaac aagtttcagt gtggctttgg  
15721 cctcaccatc ggctgagccc tcctggcccc cgccttccac gcccttccga ttccacctcc  
25 15781 acctccacct cccctgccca cagaggggag acctgagccc ccctcccttc cctccccct  
15841 tgggggtcgg gggggacatt ggaaaggagg gaccccgcca cccagcagc tgaggagggg  
15901 attctggaac tgaatggcgc ttcgggattc tgagtagcag gggcagcatg cccagtgggc  
15961 ctgggggtccc gggagggatt ccggaattga ggggcacgca ggattctgag caccaggggc  
16021 agaggcggcc agacaacctc agggaggagt gtcctggcgt cccatcctc caaagggcct  
30 16081 gggcccgcgc cgagggggca gcgagaggag cttccccatc cccggtcagt ccaccctgcc  
16141 ccgtccactt tcccatctcc tcggtataaa tcatgtttat aagttatgga agaaccggga  
16201 cattttacag aaaaaaaca aaaaacaaca aaaaatatac gtgggaaaaa aaacgatggg  
16261 aggcctccgt tttctcaagt gtgtctggcc tgttttgagc atttcatccg gagtctggcc  
16321 gccctgacct tccccagcc gcctgcaggg ggcgcagag ggccggagca cggaaagcag  
35 16381 cggatccttg atgctgcctt aagtccggct cagaggggag cagcgtggcc tggggtcgct  
16441 atcttcccat ccggaacatc tgccctgctg ggggacacta cgggccttcc cttgcctgag

nt16541 \*

16501 ggtaggggtct caaggtcact tgccccccagc ttgacctggc ggagtgggt atagaggact  
 16561 ttgtccctgc agactgcagc agcagagatg acactgtctc tgagtgcaga gatgggggca  
 16621 gggagctggg agaggggtca agctactgga acagcttcag aacaactagg gtactaggaa  
 5 16681 ctgctgtgtc agggagaagg ggctcaagga ctgcgaggcc tgggaggagg ggcctaggcc

nt16747 \*

16741 agccat gga gttgggtcac ctgtgtctga ggacttgggtg ctgtctggat ttgccaacc  
 16801 tagggctggg gtcagctgat gccaccacg actcccgagc ctccaggaaac tgaaaccctg  
 16861 tctgccccca gggctctgggg aaggaggctg ctgagtagaa ccaacccag gttaccaacc

10

nt16965 \*

16921 ccacctcagc cacccttgc cagccaaagc aaacaggccc ggcc ggcac tgggggttcc

nt17030 \*

16981 ttctcgaacc aggagttcag cctcccctga cccgcagaat cttctgatc caccgctcc

nt17098 \*

17041 aggagccagg aatgagtccc agtctctccc agttctcact gtgtgggtttt gccattc tc  
 17101 ttgctgctga accacgggtt tctcctctga aacatctggg atttataaca gggcttagga  
 17161 aagtgcagc gtctgagcgt tcaactgtggc ctgtccattg ctagccctaa cataggaccg  
 17221 ctgtgtgcca gggctgtcct ccatgctcaa tacacgttag cttgtcacca aacatacccg  
 17281 tgccgctgct ttcccagtct gatgagcaaa ggaacttgat gctcagagag gacaagtcac

nt17387 \*

17341 ttgcccgaagg tcacacagct ggcaactggc agagccagga ttcacg cct ggcaatttga  
 17401 ctccagaatc ctaaccttaa ccagaagca cggcttcaag cccctggaaa ccacaatacc  
 17461 tgtggcagcc agggggagggt gctggaatct catttcacat gtggggaggg ggcctccctg  
 17521 tgctcaaggt cacaaccaa gaggaagctg tgattaaaac ccagggtcca tttgcaaagc  
 25 17581 ctgcactttt agcaggtgca tcatactgtt cccaccctc ccatccact tctgtccagc  
 17641 cgcctagccc cactttcttt tttttctttt tttgagacag tctccctctt gctgaggctg  
 17701 gagtgcagtg gcgagatctc ggctcactgt aacctccgcc tcccgggttc aagcgattct

nt17785 \*

17761 cctgcctcag cctcccaagt agct ggatt acaggcgccc gccaccacgc ctggctaact

30

nt17874 \*

17821 tttgtatttt tagtagagat ggggtttcac catgttggcc aggctgggtc caa ctctg

nt17937 \*

17881 accttaagtg attcgccac tgtggcctcc caaagtgtg ggattacagg cgtgac acc  
 17941 gccccagcc cctcccatcc cacttctgtc cagcccccta gccctacttt ctttctggga  
 35 18001 tccaggagtc cagatcccca gcccctctc cagattacat tcattccaggc acaggaaagg  
 18061 acagggtcag gaaaggagga ctctgggcgg cagcctccac attccccttc cagcgttggc

nt18145 \*

18121 cccagaaatg gaggaggggtg tctg attac tgggagaggt gtcctccctt cctggggact  
 18181 gtgggggggtg gtcaaaagac ctctatgccc cacctccttc ctccctctgc cctgctgtgc  
 18241 ctgggggcagg gggagaacag cccacctcgt gactgggggc tggcccagcc cgccctatcc  
 5 18301 ctggggggagg gggcgggaca gggggagccc tataattgga caagtctggg atccttgagt  
 18361 cctactcagc **CCCAGCGGAG GTGAAGGACG TCCTTCCCCA GGAGCCG**gtg agaagcgcag

nt18476 \*

18421 tcggggggcac ggggatgagc tcaggggcct ctagaaagag ctgggaccct gggaa ccct  
 18481 ggcctccagg tagtctcagg agagctactc ggggtcgggc ttggggagag gaggagcggg  
 10 18541 ggtgaggcaa gcagcagggg actggacctg ggaagggctg ggcagcagag acgacccgac  
 18601 ccgctagaag gtgggggtggg gagagcagct ggactgggat gtaagccata gcaggactcc  
 18661 acgagttgtc actatcattt atcgagcacc tactgggtgt cccagtgte ctcagatctc  
 18721 cataactggg gagccagggg cagcgacacg gtagctagcc gtcgattgga gaactttaaa  
 18781 atgaggactg aattagctca taaatggaac acggcgctta actgtgaggt tggagcttag  
 18841 aatgtgaagg gagaatgagg aatgcgagac tgggactgag atggaaccgg cgggtggggag  
 18901 ggggtggggg gatggaattt gaaccccggg agaggaagat ggaattttct atggaggccg  
 18961 acctggggat ggggagataa gagaagacca ggagggagtt aaatagggaa tgggttgggg  
 19021 gcggcttggg aaatgtgctg ggattaggct gttgcagata atgcaacaag gcttggaagg  
 19081 ctaacctggg gtgaggccgg gttggggccg ggctgggggt gggaggagtc ctcactggcg  
 20 19141 gttgattgac agtttctcct tccccag**ACT GGCCAATCAC AGGCAGGAAG ATGAAGGTTT**  
 19201 **TGTGGGCTGC GTTGCTGGTC ACATTCTGG CAGG**tatggg ggcggggcct gctcggttcc

nt19311 \*

19261 ccccgtcct cccctctca tctcacctc aacctcctgg cccattcag cagacctg  
 19321 ggccccctct tctgaggctt ctgtgctgct tctggctct gaacagcgat ttgacgtct  
 25 19381 ctgggcctcg gtttccccca tcttgagat aggagttaga agttgttttg ttgttgttgt  
 19441 ttgttgttgt tgttttgttt ttttgagatg aagtctcgct ctgtcgcca ggctggagtg  
 19501 cagtggcggg atctcggtc actgcaagct ccgcctcca ggtccacgcc attctcctgc  
 19561 ctcagcctcc caagtagctg ggactacagg cacatgccac cacacccgac taactttttt  
 19621 gtattttcag tagagacggg gtttcacat gttggccagg ctggtctgga actcctgacc  
 30 19681 tcaggtgatc tgcccgtttc gatctcccaa agtgctggga ttacaggcgt gagccaccgc  
 19741 acctggctgg gagttagagg tttctaagc attgcaggca gatagtgaat accagacag  
 19801 gggcagctgt gatctttatt ctccatcacc cccacacagc cctgcctggg gcacacaagg  
 19861 aactcaata catgcttttc cgctgggcgc ggtggctcac cctgtaate ccagcacttt  
 19921 gggaggccaa ggtgggagga tcaattgagc ccaggagttc aacaccagcc tgggcaacat  
 35 19981 agtgagaccc tgtctctact aaaaatacaa aaattagcca ggcattggtgc cacacacctg  
 20041 tgctctcagc tactcaggag gctgaggcag gaggatcgct tgagcccaga aggtcaaggt



20101 tgcagtgaac catgttcagg ccgctgcact ccagcctggg tgacagagca agaccctgtt  
 20161 tataaataca taatgctttc caagtgatta aaccgactcc cccctcaccg tgcccaccat  
 20221 ggctccaaag aagcatttgt ggagcacctt ctgtgtgccc ctaggtacta gatgcctgga

nt20334 (A18T) \*

5 20281 cggggtcaga aggaccctga cccaccttga acttggttcca cacagg**ATGC CAG CCAAGG**  
 20341 **TGGAGCAAGC GGTGGAGACA GAGCCGGAGC CCGAGCTGCG CCAGCAGACC GAGTGGCAGA**  
 20401 **GCGGCCAGCG CTGGGAAC TG CACTGGGTC GCTTTTGGGA TTACCTGCGC TGGGTGCAGA**  
 20461 **CACTGTCTGA GCAGGTGCAG GAGGAGCTGC TCAGCTCCCA GGTCACCCAG GAACTGAGGt**  
 20521 gagtgtcccc atcctggccc ttgacctcc tgggtggcgg ctatacctcc ccaggtccag  
 10 20581 gtttcattct gcccctgtcg ctaagtcttg gggggcctgg gtctctgtctg gttctagctt  
 20641 cctcttccca tttctgactc ctggcttttag ctctctggaa ttctctctct cagctttgtc  
 20701 tctctctctt ccttctgac tcagtctctc aactcgtcc tggctctgtc tctgtccttc  
 20761 cctagctctt ttatatagag acagagagat ggggtctcac tgtgttgccc aggtcgtgtc  
 20821 tgaacttctg ggctcaagcg atcctcccgc ctggcctcc caaagtgtg ggattagagg  
 20881 catgagccac cttgcccggc ctctagctc cttctctgtc tctgctctg cctctgcat  
 20941 ctgctctctg catctgtctc tgtctccttc tctcggtc tgcctcgttc cttctctccc  
 21001 tcttgggtct ctctggctca tccccatctc gcccgccca tcccagcct tctccccgc  
 21061 tcccactgtg cgacaccctc ccgcctctc ggccgcagg**G CGCTGATGGA CGAGACCATG**  
 21121 **AAGGAGTTGA AGGCCTACAA ATCGGAAC TG GAGGAACAAC TGACCCCGGT GCGGAGGAG**  
 21181 **ACGCGGGCAC GGCTGTCCAA GGAGCTGCAG GCGGCGCAGG CCCGGCTGGG CGCGGACATG**  
 nt21250 (C130R)  
 21241 **GAGGACGTG GCGGCCGCCT GGTGCAGTAC CGCGGCGAGG TGCAGGCCAT GCTCGGCCAG**  
 nt21349 (R163C)  
 21301 **AGCACCGAGG AGCTGCGGGT GCGCCTCGCC TCCCACCTGC GCAAGCTG G TAAGCGGCTC**  
 25 nt21388 (R176C)  
 21361 **CTCCGCGATG CCGATGACCT GCAGAAG GC CTGGCAGTGT ACCAGGCCGG GGCCCGCGAG**  
 21421 **GGCGCCGAGC GCGGCTCAG CGCCATCCGC GAGCGCTGG GGCCCTGGT GGAACAGGGC**  
 21481 **CGCGTGCGGG CCGCCACTGT GGGCTCCCTG GCCGGCCAGC CGCTACAGGA GCGGGCCAG**  
 21541 **GCCTGGGGCG AGCGGCTGCG CGCGCGGATG GAGGAGATGG GCAGCCGGAC CCGCGACCGC**  
 30 21601 **CTGGACGAGG TGAAGGAGCA GGTGGCGGAG GTGCGCGCCA AGCTGGAGGA GCAGGCCAG**  
 21661 **CAGATACGCC TGCAGGCCGA GGCCTTCCAG GCCCGCTCA AGAGCTGGTT CGAGCCCCTG**  
 21721 **GTGGAAGACA TGCAGCGCCA GTGGGCCGGG CTGGTGGAGA AGGTGCAGGC TGCCGTGGGC**  
 21781 **ACCAGCGCCG CCCCTGTGCC CAGCGACAAT CACTGAACGC CGAAGCCTGC AGCCATGCGA**  
 21841 **CCCCACGCCA CCCCGTGCCT CCTGCCTCCG CGCAGCCTGC AGCGGGAGAC CCTGTCCCCG**  
 35 21901 **CCCCAGCCGT CCTCTGGGG TGGACCCTAG TTTAATAAAG ATTACCAAG TTTACGcat**  
 21961 ctgctggcct ccccctgtga tttctctaa gcccagcct cagtttctct tctgtccac

5

10

0052020 "B202650

25

30

35

22021 atactggcca cacaattctc agccccctcc tctccatctg tgtctgtgtg tatctttctc  
 22081 tctgcccttt tttttttttt tagacggagt ctggctctgt caccaggct agagtgcagt  
 22141 ggcacgatct tggctcactg caacctctgc ctcttgggtt caagcgattc tgctgcctca  
 22201 gtagctggga ttacaggctc acaccaccac acccggttaa tttttgtatt tttagtagag  
 22261 acgagctttc accatgttgg ccaggcaggt ctcaaactcc tgaccaagtg atccaccgcg  
 22321 cggcctccca aagtgtctgag attacaggcc tgagccacca tgcccggcct ctgccctctc  
 22381 ttctttttta gggggcaggg aaaggtctca ccctgtcacc cgccatcaca gctcactgca  
 22441 gcctccacct cctggactca agtgataagt gatcctcccg cctcagcctt tccagtagct  
 22501 gagactacag gcgcatacca ctaggattaa tttggggggg ggggtggtgtg tgtggagatg  
 22561 gggctctggct ttgttggcca ggctgatgtg gaattcctgg gctcaagcga tactcccacc  
 22621 ttggcctcct gagtagctga gactactggc tagcaccacc acaccagct ttttattatt  
 22681 atttgtagag acaaggtctc aatatgttgc ccaggctagt ctcaaaccct tgggctcaag  
 22741 agatcctccg ccatcggcct cccaaagtgc tgggattcca ggcattggggc tccgagcccg  
 22801 gcctgcccaa cttaataata cttgttctc agagttgcaa ctccaaatga cctgagattg  
 22861 gtgcctttat tctaagctat ttctattttt tttctgtgtg cattattctc ccccttctct  
 22921 cctccagtct tatctgatat ctgcctcctt cccaccacc ctgcaccca tcccaccct  
 22981 ctgtctctcc ctgttctcct caggagactc tggcttctctg ttttctcca cttctatctt  
 23041 ttatctctcc ctctacggg ttcttttctt tctccccggc ctgcttgttt ctccccaac  
 23101 ccccttcac tggttttctt cttctgccat tcagtttggg ttgagctctc tgcttctccg  
 23161 gttccctctg agctagctgt cccttcacc actgtgaact gggtttccct gcccaaccct  
 23221 cattctcttt ctttctttct tttttttttt tttttttttt tttttttttt gagacagagt  
 23281 cttgctctgt tgcccagcct ggagtgcagt ggtgcaatct tggttcactg caacctccac  
 23341 ttcccagatt caagcaattc tcctgcctca gcctccagag tagctgggat tacaggcgtg  
 23401 tcccaccaca cccgactaat ttttgtattt ttggtagaga caaggcttcg gcattgttgg  
 23461 ccaggcaggt ctggaactcc tgacctcaag taatctgcct gcctcacct cccaaagtgc  
 nt23524 \*  
 23521 tgg attaca ggcattgagc acctcaccg gaccatccct cattctccat ctttctctc  
 23581 agttgtgatg tctaccctc atgtttccca acaagcctac tgggtgctga atccaggctg  
 23641 ggaagagaag ggagcggctc ttctgtcgga gtctgcacca ggcccatgct gagacgagag  
 nt23707 \* nt23759 \*  
 23701 ctggcg tca gagaggggaa gcttggatgg aagcccagga gccgccggca ctctcttc c  
 nt23805 \*  
 23761 ctcccacccc ctgattctc agagacgggg aggagggttc ccac aacgg gggacaggct  
 23821 gagacttgag cttgtatctc ctgggccagc tgcaacatct gcttgtccct ctgcccattc  
 23881 tggctcctgc acaccctgaa cttggtgctt tccttggcac tgctctgac acccagctgg  
 23941 aggcagcacc cctcccctgg agatgactca ccagggtga gtgaggagg gaagggtcag

24001 tgtgtcacca ggcagggggc ctggctctgct gggcctgctg ctgattcacc gtatgtccag

BREAK

5 36601 catgcttag gagggacatt tcaaactctt tttacccta gactttccta ccataccca  
 36661 gagtatccag ccaggagggg aggggctaga gacaccagaa gtttagcagg gaggagggcg  
 36721 tagggattcg gggaatgaag ggatgggatt cagactaggg ccaggacca gggatggaga  
 36781 gaaagagatg agagtggttt gggggcttgg tgacttagag aacagagctg caggctcaga  
 36841 ggcacacagg agtttctggg ctccacctgc ccccttccaa cccctcagtt cccatcctcc  
 10 36901 agcagctgtt tgtgtgctgc ctctgaagtc cacactgaac aaacttcagc ctactcatgt  
 36961 ccctaaaatg ggcaaacatt gcaagcagca aacagcaaac acacagccct ccctgctgc  
 37021 tgaccttggg gctggggcag aggtcagaga cctctctggg cccatgccac ctccaacatc  
 37081 cactcgacct cttggaattt cgggtggagag gagcagaggt tgtcctggcg tggtttaggt  
 37141 agtgtgagag ggtccgggtt caaaaccact tgctgggtgg ggagtcgtca gtaagtggct  
 nt37237 \*

37201 atgccccgac cccgaagcct gtttcccat ctgtac atg gaaatgataa agacgcccac  
 37261 ctgatagggt ttttgtggca aataaacatt tggttttttt gttttgtttt gttttgtttt  
 37321 ttgagatgga ggtttgcctc gtcgcccagg ctggagtgcg gtgacacaat ctcatctcac  
 37381 cacaaccttc ccctgcctca gcctcccaag tagctgggat tacaagcatg tgccaccaca  
 37441 cctggctaata tttctatttt tagtagagac gggtttctcc atgttgggtca gcctcagcct  
 37501 cccaagtaac tgggattaca ggcctgtgcc accacaccg gctaattttt tctatttttg  
 37561 acagggacgg ggtttcacca tgttggtcag gctgggtctag aactcctgac ctcaaagat  
 37621 ccaccacact aggcctccca aagtgcacag attacaggcg tgggccaccg cacctggcca

25 BREAK

41821 aaaagatggt cttgtggggg aatgaaggac acaagcttgg tgggacctga gtccccaggc

41881 tggcatagag ccccttactc cctgtgt

//

30 = Polymorphisms (the polymorphic nt is numbered)

**Bold** = ApoE transcribed sequences (exons 1 - 4)

Grey shaded = Contains ApoE enhancer

Underline = Coding Region of the ApoE gene

\* = Polymorphisms not previously described in the art

35

**BEST COPY AVAILABLE**

VGX Symbol	VGX Database	VGX GenBank	GEN-CBX	GEN-CBX	GEN-CBX	GEN-P0	GEN-CBX	GEN-CBX	GEN-P0	GEN-CBX	GEN-CBX
Amino Acid Change	A(Silent)/T(Silent)	T(Silent)/C(Silent)	T(Silent)/G(Silent)	C(Silent)/G(Silent)	G(Silent)/A(Silent)	G(Alanine)/A(Threonine)	T(Cysteine)/C(Arginine)	C(Arginine)/T(Cysteine)			
1	Genotype	A	T	TG	G	G	[TC]	C	C		
	Haplotype 1	A	T	G	G	G	T	C	C		
	Haplotype 2	A	T	T	G	G	C	C	C		
3	Genotype	[AT]	T	G	G	G	T	C	C		
	Haplotype 1	A	T	G	G	G	T	C	C		
	Haplotype 2	T	T	G	G	G	T	C	C		
5	Genotype	[AT]	T	G	G	G	T	C	C		
	Haplotype 1	T	T	G	G	G	T	C	C		
	Haplotype 2	A	T	G	G	G	T	C	C		
6	Genotype	[AT]	T	G	G	G	T	C	C		
	Haplotype 1	A	T	G	G	G	T	C	C		
	Haplotype 2	T	T	G	G	G	T	C	C		
7	Genotype	A	T	G	G	G	T	C	C		
	Haplotype 1	A	T	G	G	G	T	C	C		
	Haplotype 2	A	T	G	G	G	T	C	C		
8	Genotype	A	T	T	C	G	T	C	C		
	Haplotype 1	A	T	T	C	G	T	C	C		
	Haplotype 2	A	T	T	C	G	T	C	C		
11	Genotype	[AT]	T	G	G	G	T	C	C		
	Haplotype 1	A	T	G	G	G	T	C	C		
	Haplotype 2	T	T	G	G	G	T	C	C		
12	Genotype	A	T	TG	G	G	T	C	C		
	Haplotype 1	A	T	T	G	G	T	C	C		
	Haplotype 2	A	T	G	G	G	T	C	C		
13	Genotype	A	T	TG	G	G	[TC]	C	C		
	Haplotype 1	A	T	G	G	G	T	C	C		
	Haplotype 2	A	T	T	G	A	C	C	C		
14	Genotype	A	T	TG	[CG]	G	T	C	C		
	Haplotype 1	A	T	G	G	G	T	C	C		
	Haplotype 2	A	T	T	C	C	T	C	C		
15	Genotype	A	[TC]	TG	[CG]	G	T	C	[CT]		
	Haplotype 1	A	C	G	G	G	T	C	T		
	Haplotype 2	A	T	T	C	G	T	C	C		
16	Genotype	A	T	T	C	G	T	C	C		
	Haplotype 1	A	T	T	C	G	T	C	C		
	Haplotype 2	A	T	T	C	G	T	C	C		
17	Genotype	A	T	TG	[CG]	G	T	C	C		
	Haplotype 1	A	T	G	G	G	T	C	C		
	Haplotype 2	A	T	T	C	G	T	C	C		
18	Genotype	A	T	T	C	G	[GA]	C	C		
	Haplotype 1	A	T	T	C	G	T	C	C		
	Haplotype 2	A	T	T	C	G	T	C	C		
19	Genotype	A	T	T	C	G	T	C	C		
	Haplotype 1	A	T	T	C	G	T	C	C		
	Haplotype 2	A	T	T	C	G	T	C	C		

# 005207 Table 2026960

WV/P#	VGX Symbol	VGX Database	GenBank	Amino Acid Change	GEN-CBX 1494 17874	GEN-CBX 1557 17937	GEN-CBX 1765 18145	GEN-CBX 2096 18476	GEN-CBX 29311 19311	GEN-P0 112 20334	GEN-P0 448 21250	GEN-CBX 4969 21349	GEN-CBX 5008 21388
20	Genotype	A	T	[TG]	[CG]	G	T	C	T	C	C	C	C
	Haplotype 1	A	T	G	G	G	T	G	T	G	C	C	C
	Haplotype 2	A	T	T	C	G	T	C	T	G	C	C	C
21	Genotype	A	T	[TG]	[CG]	G	T	C	T	C	C	C	C
	Haplotype 1	A	T	G	G	G	T	G	T	G	C	C	C
	Haplotype 2	A	T	T	C	G	T	C	T	G	C	C	C
24	Genotype	A	T	[TG]	[CG]	G	T	C	T	C	C	C	C
	Haplotype 1	A	T	G	G	G	T	G	T	G	C	C	C
	Haplotype 2	A	T	T	C	G	T	C	T	G	C	C	C
25	Genotype	A	T	G	G	G	T	G	T	G	C	C	C
	Haplotype 1	A	T	G	G	G	T	G	T	G	C	C	C
	Haplotype 2	A	T	G	G	G	T	G	T	G	C	C	C
26	Genotype	A	T	T	C	G	T	C	T	C	C	C	C
	Haplotype 1	A	T	T	C	G	T	C	T	C	C	C	C
	Haplotype 2	A	T	T	C	G	T	C	T	C	C	C	C
27	Genotype	A	T	[TG]	[CG]	G	T	C	T	C	C	C	C
	Haplotype 1	A	T	G	G	G	T	G	T	G	C	C	C
	Haplotype 2	A	T	T	C	G	T	C	T	G	C	C	C
28	Genotype	A	C	G	G	G	T	G	T	G	C	C	C
	Haplotype 1	A	C	G	G	G	T	G	T	G	C	C	C
	Haplotype 2	A	C	G	G	G	T	G	T	G	C	C	C
29	Genotype	A	T	G	G	G	T	G	T	G	C	C	C
	Haplotype 1	A	T	G	G	G	T	G	T	G	C	C	C
	Haplotype 2	A	T	G	G	G	T	G	T	G	C	C	C
30	Genotype	A	T	[TG]	C	G	T	C	T	C	C	C	C
	Haplotype 1	A	T	T	C	G	T	C	T	C	C	C	C
	Haplotype 2	A	T	G	C	G	T	C	T	C	C	C	C
31	Genotype	A	T	[TG]	[CG]	G	T	C	T	C	C	C	C
	Haplotype 1	A	T	G	G	G	T	G	T	G	C	C	C
	Haplotype 2	A	T	T	C	G	T	C	T	G	C	C	C
32	Genotype	A	T	[TG]	C	G	T	C	T	C	C	C	C
	Haplotype 1	A	T	T	C	G	T	C	T	C	C	C	C
	Haplotype 2	A	T	G	C	G	T	C	T	C	C	C	C
33	Genotype	[AT]	T	[TG]	[CG]	G	T	C	T	C	C	C	C
	Haplotype 1	T	T	T	C	G	T	G	T	G	C	C	C
	Haplotype 2	A	T	G	G	G	T	C	T	G	C	C	C
34	Genotype	A	T	[TG]	G	[GA]	T	C	T	C	C	C	C
	Haplotype 1	A	T	G	G	G	T	G	T	G	C	C	C
	Haplotype 2	A	T	T	G	A	T	C	T	C	C	C	C
35	Genotype	A	[TC]	[TG]	[CG]	G	T	C	T	C	C	C	[CT]
	Haplotype 1	A	C	G	G	G	T	G	T	G	C	C	C
	Haplotype 2	A	T	T	C	G	T	C	T	G	C	C	C
36	Genotype	A	[TC]	[TG]	[CG]	G	T	C	T	C	C	C	C
	Haplotype 1	A	T	G	G	G	T	G	T	G	C	C	C
	Haplotype 2	A	C	T	C	G	T	C	T	G	C	C	C

WVPH	VGX Symbol	GEN-CBX	GEN-CBX	GEN-CBX	GEN-CBX	GEN-P0	GEN-CBX	GEN-CBX
	VGX Database	1494	1557	1765	2096	29311	112	5008
	GenBank	17874	17937	18145	18476	19311	20334	21388
	Amino Acid Change	A(Silent)/T(Silent)	T(Silent)/C(Silent)	T(Silent)/G(Silent)	C(Silent)/G(Silent)	G(Silent)/A(Silent)	G(Alanine)/A(Threonine)	C(Arginine)/T(Cysteine)
38	Genotype	[AT]	T	G	G	G	T	C
	Haplotype 1	A	T	G	G	G	T	C
	Haplotype 2	T	T	G	G	G	T	C
39	Genotype	[AT]	[TC]	[TG]	[CG]	G	T	[CT]
	Haplotype 1	T	T	T	C	G	T	C
	Haplotype 2	A	C	G	G	G	T	T
40	Genotype	A	T	[TG]	G	G	T	C
	Haplotype 1	A	T	T	G	G	T	C
	Haplotype 2	A	T	G	G	G	T	C
41	Genotype	[AT]	T	G	[CG]	G	[TC]	C
	Haplotype 1	T	T	G	C	G	T	C
	Haplotype 2	A	T	G	G	G	C	C
42	Genotype	A	[TC]	[TG]	[CG]	G	T	C
	Haplotype 1	A	T	G	G	G	T	C
	Haplotype 2	A	C	T	C	G	T	C
44	Genotype	A	C	T	C	G	T	C
	Haplotype 1	A	C	T	C	G	T	C
	Haplotype 2	A	C	T	C	G	T	C
45	Genotype	A	T	T	[CG]	[GA]	[TC]	C
	Haplotype 1	A	T	T	C	G	T	C
	Haplotype 2	A	T	T	G	A	C	C
46	Genotype	[AT]	T	G	G	G	T	C
	Haplotype 1	A	T	G	G	G	T	C
	Haplotype 2	T	T	G	G	G	T	C
47	Genotype	[AT]	T	[TG]	G	G	[TC]	C
	Haplotype 1	T	T	T	G	G	C	C
	Haplotype 2	A	T	G	G	G	T	C
48	Genotype	[AT]	[TC]	[TG]	[CG]	G	T	[CT]
	Haplotype 1	T	T	T	C	G	T	C
	Haplotype 2	A	C	G	G	G	T	T
49	Genotype	A	T	[TG]	[CG]	G	T	C
	Haplotype 1	A	T	G	G	G	T	C
	Haplotype 2	A	T	T	C	G	T	C
50	Genotype	A	T	T	[CG]	[GA]	[TC]	C
	Haplotype 1	A	T	T	C	G	T	C
	Haplotype 2	A	T	T	G	A	C	C
51	Genotype	A	[TC]	[TG]	G	[GA]	[TC]	[CT]
	Haplotype 1	A	C	T	G	G	T	T
	Haplotype 2	A	T	T	G	A	C	C
52	Genotype	[AT]	T	G	G	G	T	C
	Haplotype 1	A	T	T	G	G	T	C
	Haplotype 2	T	T	G	G	G	T	C
54	Genotype	A	[TC]	[TG]	[CG]	G	T	C
	Haplotype 1	A	T	G	G	G	T	C
	Haplotype 2	A	C	T	C	G	T	C

WWP#	VGX Symbol	GEN-CBX	GEN-CBX	GEN-CBX	GEN-CBX	GEN-P0	GEN-CBX	GEN-CBX	GEN-P0	GEN-CBX	GEN-CBX
	VGX Database	1494	1557	1765	2096	112	29311	4969	448	5008	GEN-CBX
	GenBank	17874	17937	18145	18476	20334	19311	21349	21250	21388	GEN-CBX
	Amino Acid Change	A(Silent)/T(Silent)	T(Silent)/C(Silent)	T(Silent)/G(Silent)	C(Silent)/G(Silent)	G(Alanine)/A(Threonine)	G(Silent)/A(Silent)	C(Arginine)/T(Cysteine)	T(Cysteine)/C(Arginine)	C(Arginine)/T(Cysteine)	GEN-CBX
58	Genotype	[AT]	T	[TG]	[CG]	G	G	C	T	C	C
	Haplotype 1	T	T	T	C	G	G	C	T	C	C
	Haplotype 2	A	T	G	G	G	G	C	T	C	C
59	Genotype	A	T	G	G	G	G	[CT]	T	C	C
	Haplotype 1	A	T	G	G	G	G	C	T	C	C
	Haplotype 2	A	T	G	G	G	G	T	T	C	C
60	Genotype	[AT]	T	G	G	G	G	C	T	C	[CT]
	Haplotype 1	T	T	G	G	T	G	C	T	T	C
	Haplotype 2	A	T	G	G	G	G	C	T	C	C
61	Genotype	[AT]	T	[TG]	G	G	[GA]	C	[TC]	C	[CT]
	Haplotype 1	T	T	G	G	G	G	C	T	C	T
	Haplotype 2	A	T	T	G	A	A	C	C	C	C
62	Genotype	[AT]	T	G	G	G	G	C	T	C	[CT]
	Haplotype 1	T	T	G	G	G	G	C	T	T	C
	Haplotype 2	A	T	G	G	G	G	C	T	C	C
63	Genotype	A	T	T	C	G	G	C	T	C	C
	Haplotype 1	A	T	T	C	G	G	C	T	C	C
	Haplotype 2	A	T	T	C	G	G	C	T	C	C
66	Genotype	[TC]	T	G	G	G	G	C	T	C	[CT]
	Haplotype 1	A	C	G	G	G	G	C	T	T	C
	Haplotype 2	A	T	G	G	G	G	C	T	C	C
67	Genotype	A	T	[TG]	[CG]	G	G	C	T	C	C
	Haplotype 1	A	T	G	G	G	G	C	T	C	C
	Haplotype 2	A	T	T	C	G	G	C	T	C	C
68	Genotype	[AT]	T	T	[CG]	G	[GA]	C	[TC]	C	C
	Haplotype 1	T	T	T	C	G	G	C	T	C	C
	Haplotype 2	A	T	T	G	A	A	C	C	C	C
69	Genotype	A	T	[TG]	G	G	[GA]	C	[TC]	C	C
	Haplotype 1	A	T	G	G	G	G	C	T	C	C
	Haplotype 2	A	T	T	G	A	A	C	C	C	C
70	Genotype	[AT]	T	G	G	G	G	C	T	C	C
	Haplotype 1	A	T	G	G	G	G	C	T	C	C
	Haplotype 2	A	T	G	G	G	G	C	T	C	C
71	Genotype	A	T	[TG]	[CG]	G	G	C	T	C	C
	Haplotype 1	A	T	G	G	G	G	C	T	C	C
	Haplotype 2	A	T	T	C	G	G	C	T	C	C
72	Genotype	[TC]	T	[TG]	[CG]	G	G	C	T	C	[CT]
	Haplotype 1	A	C	G	G	G	G	C	T	T	C
	Haplotype 2	A	T	T	C	G	G	C	T	C	C
73	Genotype	A	T	[TG]	[CG]	G	G	C	T	C	C
	Haplotype 1	A	T	G	G	G	G	C	T	C	C
	Haplotype 2	A	T	T	C	G	G	C	T	C	C
74	Genotype	[AT]	T	[TG]	[CG]	G	G	C	T	C	C
	Haplotype 1	T	T	G	G	G	G	C	T	C	C
	Haplotype 2	A	T	T	C	G	G	C	T	C	C

**FOR OFFICIAL USE ONLY**

WWP#	VGX Symbol	GEN-CBX	GEN-CBX	GEN-CBX	GEN-CBX	GEN-P0	GEN-CBX	GEN-CBX	GEN-CBX
	VGX Database	1494	17874	1557	17937	1765	18145	2096	18476
	GenBank	A(Silent)/T(Silent)	T(Silent)/C(Silent)	T(Silent)/G(Silent)	C(Silent)/G(Silent)	G(Silent)/A(Silent)	G(Alanine)/A(Threonine)	T(Cysteine)/C(Arginine)	C(Arginine)/T(Cysteine)
75	Genotype	[AT]	T	[TG]	[CG]	G	T	C	C
	Haplotype 1	T	T	G	G	G	T	C	C
	Haplotype 2	A	T	T	C	G	T	C	C
78	Genotype	[AT]	T	[TG]	[CG]	G	T	[CT]	C
	Haplotype 1	T	T	T	C	G	T	C	C
	Haplotype 2	A	T	G	G	G	T	T	C
79	Genotype	[AT]	T	T	G	G	C	C	C
	Haplotype 1	T	T	T	G	G	C	C	C
	Haplotype 2	A	T	T	G	G	C	C	C
80	Genotype	T	T	[TG]	[CG]	G	[TC]	C	C
	Haplotype 1	T	T	G	G	G	C	C	C
	Haplotype 2	T	T	T	C	G	T	C	C
81	Genotype	A	T	G	G	G	T	C	C
	Haplotype 1	A	T	G	G	G	T	C	C
	Haplotype 2	A	T	G	G	G	T	C	C
84	Genotype	A	T	T	C	G	T	C	C
	Haplotype 1	A	T	T	C	G	T	C	C
	Haplotype 2	A	T	T	C	G	T	C	C
93	Genotype	A	T	T	C	G	T	C	C
	Haplotype 1	A	T	T	C	G	T	C	C
	Haplotype 2	A	T	T	C	G	T	C	C
95	Genotype	A	T	[TG]	[CG]	G	T	C	C
	Haplotype 1	A	T	G	G	G	T	C	C
	Haplotype 2	A	T	T	C	G	T	C	C
101	Genotype	A	T	[TG]	[CG]	G	T	C	C
	Haplotype 1	A	T	G	G	G	T	C	C
	Haplotype 2	A	T	T	C	G	T	C	C
102	Genotype	A	T	T	C	G	T	C	C
	Haplotype 1	A	T	T	C	G	T	C	C
	Haplotype 2	A	T	T	C	G	T	C	C
109	Genotype	[AT]	T	G	G	G	T	C	C
	Haplotype 1	A	T	G	G	G	T	C	C
	Haplotype 2	T	T	G	G	G	T	C	C
110	Genotype	A	T	[TG]	[CG]	G	T	C	C
	Haplotype 1	A	T	G	G	G	T	C	C
	Haplotype 2	A	T	T	C	G	T	C	C
111	Genotype	A	T	G	G	G	T	C	C
	Haplotype 1	A	T	G	G	G	T	C	C
	Haplotype 2	A	T	G	G	G	T	C	C
112	Genotype	A	T	G	G	G	T	C	C
	Haplotype 1	A	T	G	G	G	T	C	C
	Haplotype 2	A	T	G	G	G	T	C	C
113	Genotype	[AT]	T	[TG]	[CG]	G	T	C	C
	Haplotype 1	T	T	T	C	G	T	C	C
	Haplotype 2	A	T	G	G	G	T	C	C





Table 4. ApoE haplotypes

#	16541	16747	16965	17030	17098	17387	17785	17874	17937	18145	18476	19311	20334	21250	21349	21388	23524	23707	23759	23805
1	C	G	C	C	G	C	A	A	T	T	C	G	G	T	C	C	G	C	T	C
2	C	G	C	C	G	C	A	A	T	T	C	G	A	T	C	C	G	C	T	C
3	C	G	C	C	G	C	A	A	T	T	C	G	G	T	C	C	G	C	C	C
4	C	G	C	C	G	C	A	A	T		C	G	G	T	C	C	A	C	T	C
5	C	G	C	C	G	C	A	A	T		C	G	G	T	T	C	G	C	T	C
6	C	G	C	C	G	C	A	A	T	G	C	G	G	T	C	C	G	C	T	C
7	C	G	C	C	A	C	A	A	T		G	G	G	T	C	C	G	C	T	C
8	C	G	C	C	A	C	A	A	T	G	G	G	G	T	C	C	G	C	T	C
9	C	G	C	C	A	C	A	A	T	G	G	G	G	T	C	C	G	C	C	C
10	C	G	T	C	A	C	A	A	T	G	G	G	G	T	C	C	G	C	C	C
11	C	G	C	C	A	C	A	A	T	G	G	G	G	T	C	C	G	C	T	C
12	C	T	C	G	G	C	G	A	T		G	G	G	C	C	C	G	C	C	C
13	C	T	C	G	G	C	G	A	T		G	G	A	C	C	C	G	C	C	C
14	C	G	T	C	A	C	A		T	G	G	G	G	T	C	C	G	C	T	C
15	C	G	C	C		C	A	A	T		G	G	G	T	C	C	G	C	C	C
16	C	G	C	C	G	C	A	A	T		G	G	G	T	C	C	G	C		C
17	C	G	C	C		C	A	A	T	G	G	G	G	T	C		G	A	C	C
18	C	G	C	C		C	A	A	T	G	G	G	G	T	C	T	G		C	C
19	T	G	C	C		C	A		T		G	G	G	T	C	C	G	C	C	C
20	C		C	G	G	C		A	T			G	G		C	C	G	C		C
21	C	T	C		G	C		A	T			G	G		C	C	G	C		C
22	C		C	G	G	C		A	T			G	G		C	C	G	C		C
23	C	G	C	C		C	A	A	T			G	G	T	C	C	G	C		G
24	C	G	C	C		T	A		T	G	G	G	G		C	C	G	C	C	C
25	C	G	C	C			A	T	T	G	G	G	G		C	C	G	C	C	C
26	C	G	T	C	A	C	A		T	G	G	G		T	C	C	G	C		C
27	C	G		C	A	C	A	T	T	G	G	G		T	C	C	G	C		C
28	C	G		C	A	C	A	T	T	G	G	G		T	C	C	G	C	T	C
29	C	G	C	C	G	C	A	A	T	T		A	G		C	C	G	C		C
30	C	G	C	C	G	C	A	A	T	T			G	C	C	C	G	C		C
31	C	G	C	C	G	C	A	T	T		G		G		C	T	G		C	C
32	C	G	C	C	G	C	A		T	T	G		G		C	T	G		C	C
33	C	G	C	C	G	C	A		T		G	A	G		C	T	G		C	C
34	C	G	C	C	G	C	A		T		G		G	C	C	T	G		C	C
35	C	G	C	C	G	C	A		T		G		G		C	T	G	A	C	C
36	C	G	C	C	G	T	A		T			G	G	T		C	G	C		C
37	C	G	C	C	G		A	T	T			G	G	T		C	G	C		C
38	C	G	C	C		C	A	A	C			G	G	T	C		G			C
39	C	G	C	C		C	A	A				G	G	T	C		G	A		C
40	C	G	C	C	A	C	A	A				G	G	T	C		G			C
41	C	G	C	C		C	A		C			G	G	T	C		G			C
42	C	G	C	C		C	A					G	G	T	C	T	G			C

005201 " 32046960

Table 5. One useful group of ApoE haplotypes.

#	16747	17030	17785	19311	23707	GENOTYPE
1	G	C	A	G	C	E3-like
2	T	G	G	G	C	E4-like
3	G	C	A	A	C	E4-like
4	G	C	A	G	A	E2-like
5			A	G	C	E3-like
6			G	G	C	E4-like
7			A	A	C	E4-like
8			A	G	A	E2-like
9		C		G	C	E3-like
10		G		G	C	E4-like
11		C		A	C	E4-like
12		C		G	A	E2-like
13	G			G	C	E3-like
14	T			G	C	E4-like
15	G			A	C	E4-like
16	G			G	A	E2-like
17		C	A	G	C	E3-like
18		G	G	G	C	E4-like
19		C	A	A	C	E4-like
20		C	A	G	A	E2-like
21	G	C		G	C	E3-like
22	T	G		G	C	E4-like
23	G	C		A	C	E4-like
24	G	C		G	A	E2-like
25	G		A	G	C	E3-like
26	T		G	G	C	E4-like
27	G		A	A	C	E4-like
28	G		A	G	A	E2-like

005201" 82026960